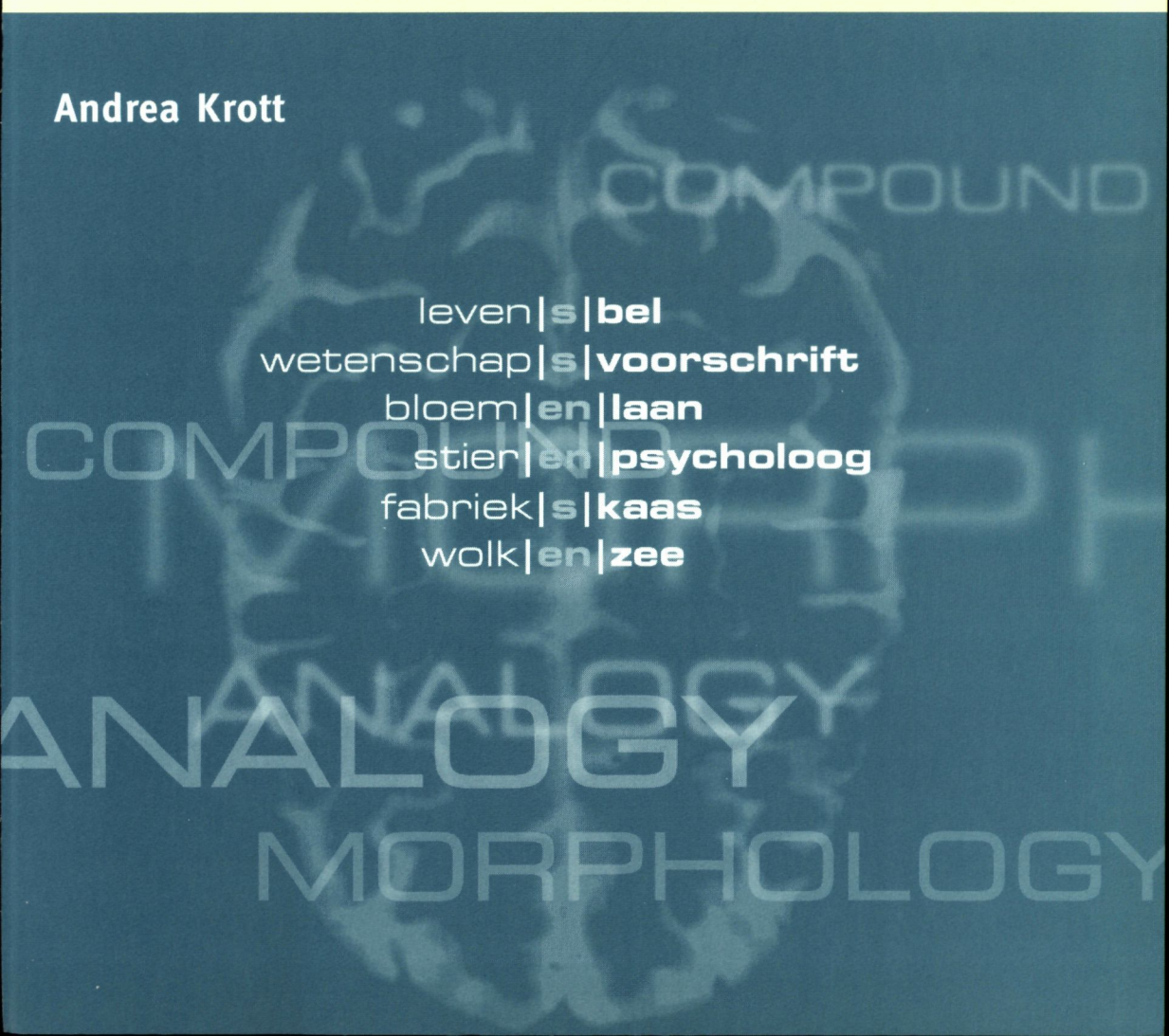
MPI SERIES
IN PSYCHOLINGUISTICS

ANALOGY IN MORPHOLOGY

THE SELECTION OF LINKING ELEMENTS IN DUTCH COMPOUNDS

Andrea Krott



Analogy in Morphology

The Selection of Linking Elements in Dutch Compounds

ISBN: 90-76203-11-3
Cover design: Linda van den Akker, Inge Doehring
Cover illustration: Marcus Krott
Printed and bound by Ponsen & Looijen bv, Wageningen

Analogy in Morphology

The Selection of Linking Elements in Dutch Compounds

een wetenschappelijke proeve
op het gebied van Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op donderdag 20 december 2001,
des namiddags om 3.30 uur precies

door
Andrea Krott
geboren op 4 augustus 1969 te Aachen (Duitsland)

Promotor:	Prof. dr. R. Schreuder
Co-promotor:	Dr. R. H. Baayen
Manuscriptcommissie:	Prof. dr. W. Vonk
	Prof. dr. A. Neijt
	Prof. dr. W. Daelemans (University of Antwerpen, Belgium)

The research reported in this thesis was supported by a PIONIER grant from the Dutch National Research Council NWO, the Faculty of Arts of the University of Nijmegen (The Netherlands) and the Max-Planck-Institut für Psycholinguistik, Nijmegen (The Netherlands).

Für meine Eltern
Sotaro-ni sasageru

Acknowledgments

Finally, the work is done and the book is finished! There is time to sit back – even if just for a moment – and to think about all the people without whom I would never have reached this point. Here I want to take the opportunity to express my gratitude for their invaluable help, support, and friendship.

First of all, I want to thank Harald Baayen. He offered me a place in his PIONIER project, a very special project with very special people. Harald is the beating heart of the project, spreading lots of enthusiasm and energy. This energy kept me moving in the last two and a half years and stopped me from hesitating in front of any upcoming hurdle. I am also deeply indebted to Rob Schreuder who is best characterized as the wise father in our small group of 'pioneers'. His experience, deliberateness, and calmness helped me to focus on the essentials. (Thanks for reminding me, Rob, that an accepted paper is a reason to celebrate!) Harald and Rob introduced me to the secrets of good research. I will never forget the sometimes long and intense but always enormously fruitful sessions of discussion, writing, and brainstorming.

Furthermore, I owe many thanks to Wolfgang Dressler for his interest in my work. He made it possible for me to run an experiment at the University of Vienna, which has resulted in an ongoing collaborative study of German linking elements. During our many discussions of this work, I gained a greater understanding of the power of the traditional approach to morphology and I learned to sharpen my arguments.

I would also like to acknowledge the members of the reading committee, Walter Daelemans, Anneke Neijt, and Wietske Vonk (in alphabetical order), for their critical and helpful comments on the manuscript. Anneke Neijt also contributed to this thesis by recruiting 200(!) students for one of my paper-and-pencil experiments. In addition, I benefited from comments and criticism that Anneke Neijt and Walter Daelemans gave on earlier drafts of the papers in this thesis.

For their practical support, I am indebted to a lot of students that helped me with running experiments. I especially want to mention Loes Krebbers, Renate Pluym,

and Kathrin Delhongue who coded the semantic features of thousands of compounds. Loes also assisted me in designing and conducting some of the experiments in chapter 4. Apart from the data coded by student assistents, I profited a lot from a list of 34,000 German compounds generously provided by Arne Fitschen and Ulrich Heid (University of Stuttgart).

Nivja de Jong, Mirjam Ernestus, Fermin Moscoso del Prado Martin, and Rachel Kemps joined the PIONIER group over the last two and a half years. Thank you all for your friendship and support! I especially enjoyed the time with Nivja, when we were the only students in the project. We used to tell each other every new result, which often ended in deep discussions. Nivja was also one of the most careful readers of my papers.

But also outside of the PIONIER group I had a lot of friends, both in the IWTS group and in the Max Planck Institute. There are too many to list them all. Let me just mention a few: Birgit Hellwig was a great friend and always listened patiently to my stories of the ups and downs of my Ph.D. project and supplied me with lots of chocolate. Ulrike Heinzel is one of my oldest friends and colleagues in this institute. I am very happy that she and Nivja de Jong agreed to be my paranimfs, supporting me on the day of my defense. Dirk Janssen and Femke van der Meulen were of big help with creating the layout of this thesis and transfering it into the right format for the printer.

My thanks are also due to the Dutch Research Council (NWO), the Faculty of Arts of the University of Nijmegen, and the Max-Planck-Institut für Psycholinguistik in Nijmegen who supported my project financially. Without them I would not have participated in so many conferences (Aix-en-Provence, Provo (Utah), Vienna, and Montreal) and I would not have run an experiment with German participants in Vienna. In the latter part of my project, I even had the luxury of getting help from student assistants who ran some of my experiments.

I am grateful for the professional support of the technical and administrative staff of the Max Planck Institute. Special thanks go to Christa Hausmann-Jamin who kept my computer up and running.

I do not want to forget the person who initiated my interest in language research: Reinhard Köhler. The vital spark of his enthusiams for research reached me while working on my master's thesis at the University of Trier.

Neben all den genannten Personen möchte ich vor allem meinen Eltern dafür danken, dass sie mich bei all meinen Plänen immer voll und ganz unterstützen. Meinem Bruder, Marcus Krott, verdanke ich den großartigen Entwurf des Buchein-

bands, auf den ich sehr stolz bin.

Saigo ni, Sotaro, itsumo watashi o gekirei shitekure, konki zuyoku shienshitekureta koto, soshite, itsumo watashi no tameni soba ni itekureta koto, kokoro kara kansha shiteimasu.

Nijmegen, July 2001.

Contents

Rules versus analogy

Every week we are exposed to new words. We hear them in conversations or we read them, for example, in newspapers. From the study of Baayen & Renouf (1996) it appears that each issue of the British newspaper *Times* from September 1989 until December 1992 contains about one novel word ending in *-ness* and more than one novel word ending in *-ly*.

What do speakers do when they construct new words? According to standard linguistic theories, speakers have two mechanisms at their disposal to form new words: rules and analogy (e.g., Anshen & Aronoff, 1988; Pinker, 1991, 1997, 1999; Pinker & Prince, 1991; Marcus, Brinkman, Clahsen, Wiese & Pinker, 1995; Clahsen, 1999). A morphological rule is assumed to be an abstract generalization of a pattern found in an usually large set of complex words, it has an explicit representation in the speaker's mind and is part of the speaker's competence. In contrast to rule-based word formation, analogical word formation is assumed to capture the construction of novel forms that are not based on general patterns but on a single word or perhaps on a small set of exemplars. In this view of analogy, speakers search their lexicons for similar words and consciously craft a new form in analogy to these words. Analogical processes and rules are understood as being cognitively distinguishable, subserved by different cortical areas.

In this standard theory, rules and analogy are understood as accounting for two types of novel words, novel regular words and novel irregular words. Novel regular words are formed by rules, novel irregular words by analogy. Thanks to the existence of rules, regular words (*walked*) do not have to be stored in memory since they can be built or decomposed on the fly whenever they are needed. Irregular words (*went*), however, are assumed to be stored in the mental lexicon and to pre-empt the use of rules (which would produce incorrect regular forms such as

**goed*). Because irregular words are stored, they may at times provide the basis for the analogical creation of a novel word. Bybee & Slobin (1982) report that participants produced, for instance, **hept* as the past tense form of *heap* in analogy to *sleep/slept*.

In the standard linguistic approach, rules are typically seen as being productive and analogy as unproductive. In other words, rules are taken to account for most new forms, while analogy is taken to be used only sporadically for rare and exceptional words. Because of the division of labor between rules and analogy, this traditional model of word formation is known as a dual-route model.

This division of labor between analogy and rules is not accepted by all, however. For instance, Bybee (1985, 1988, 1995) offers an alternative framework in which rules are extreme forms of analogy. In her model, all words and their inflectional forms are stored in the brain and connected according to their semantic, morphological, and grammatical similarities. Both highly productive rules and unproductive sub-regularities are highly reinforced representational patterns that she refers to as 'schemas'. These schemas differ from traditional rules, because they are closely tied to the forms that they represent instead of being abstract generalizations. Bybee's model has not been implemented as a computational model. However, as Bybee (1988) points out, the basic idea that patterns are built up by registering probabilities and that rules do not have to be explicitly formulated as independent mental entities that exist separately from the data, is shared by connectionist models (e.g., Rumelhart & McClelland, 1986; Plunkett & Juola, 1999; Rueckl, Mikolinski, Raveh, Miner, & Mars, 1997).

Connectionist models make use of artificial neural networks in which symbolic units like words or morphemes are typically replaced by distributed representations. They are single-route models, since one and the same network is used to model the formation of both regular and irregular forms. A common network architecture consists of a set of input units, a set of output units, and an intermediate set of hidden units that, together with the weighted links between the units, provide the means for nonlinear classification. Such models may reach an acceptable level of classification performance that resembles the behavior of speakers in forming words when extensively trained on a large set of instances with constant adjustment of the weights on the connections. Plunkett & Juola (1999), for example, model the past tense of English verbs. Their input units represent the sound of the uninflected form, their output units the sound of the appropriate past tense form. After training, the model creates the past tense form of both regular and irregular verbs.

There are also non-connectionist single-route models: the Analogical Model of Language (AML) developed by Skousen (1989) and a large range of memory-based algorithms combined in the Tilburg Memory Based Learner (TiMBL) by Daelemans, Zavrel, van der Sloot & van den Bosch (2000). In this thesis, I will use the term TiMBL to refer to its algorithms. Both in AML and TiMBL, all regular and irregular wordforms are stored as separate symbolic units, representing previous experience. With the help of similarity measures, that are defined over user-specified features, a target word is compared with the exemplars in an instance base. The set of exemplars that are most similar to the target serves as the analogical set (for details of the algorithms, see chapters 2 and 3.).

To date, studies that address the question whether explicit rules do exist and whether they are necessary to explain speakers' behavior, have focused mainly on inflection, in particular on the formation of past tense forms of English verbs (e.g., Rumelhart & McClelland, 1986; Plunkett & Juola, 1999; Pinker & Prince, 1991; Marcus, Brinkman, Clahsen, Wiese & Pinker, 1995; Clahsen, 1999). Recently, derivational word formation has been included into the discussion by Rueckl et al. (1997) and by Seidenberg and Gonnerman (2000). This thesis presents a new testing ground, the selection of linking elements for novel Dutch compounds. This process is fully productive, but, interestingly, it resists analyses in terms of rules.

In order to understand the phenomenon of Dutch linking elements, let us first look at the predictability of linking elements in compounds in general and at linking elements in Dutch compounds in particular.

Linking elements

Linking elements in compounds occur in various languages across different language families. They are also referred to as interfixes (Dressler, Libben, Stark, Pons & Jarema, 2001), juncture suffixes (Plank, 1976), connectives, linking phonemes, or linking morphemes (Schreuder, Neijt, van der Weide, & Baayen, 1998; Kehayia, Jarema, Tsapkini, Perlak, Ralli & Kadzielawa, 1999). In the light of predictability, linking elements form an inhomogeneous phenomenon. In English, a linking *-s-* can be found in frozen forms such as *grand+s+manship*, *hunt+s+man*, *state+s+man*, or *lamb+s+wool*.[1] These forms are not predictable and they have to be stored in the

[1]This linker has to be distinguished from the real plural suffix *-s-* that can also appear in compounds as in *parks commissioner*, *sales receipt*, *buildings inspector* etc. The linking *-s-* is, in contrast

lexicon. In other languages, however, linking elements are either fully predictable or partly predictable. Let us have a look at examples of both possibilities.

Languages with fully predictable linking elements are, for instance, Russian and Zoque, a Mix-Zoque language spoken in Mexico. Russian root-root compounds contain *-o-* when the first root ends in a soft consonant as in *par-o-voz* (steam-O-carry 'locomotive'), otherwise they contain *-e-* as in *pyl-e-sos* (dust-E-suck 'vacuum cleaner') (Unbegaun, 1967). In Zoque, nominal compounds occur with a connective vowel that is determined by vowel spreading ([kuhj]+[V]+[aj] 'tree+V+leave' > *kuju'aj* 'trees leaf') (Herrera, 1995). The fully predictable linking elements of Russian and Zoque are easily accounted for in terms of rules.

Linking phenomena in compounds that are partly predictable can be found in Germanic languages such as German, Danish, Dutch, Afrikaans, Swedish, and Norwegian. An example of a detailed attempt to predict the linking elements *-s-* and *-e-* in Afrikaans (e.g., *skeep+s+maat* lit. ship+S+mate 'shipmate'; *lip+e+taal* lit. lip+E+language 'lip language') is Botha (1968). He tries to analyze the distribution of linking elements in a generative framework. After searching for systematicity in the phonological context of the linkers, he concludes that their use is not predictable and all compounds have to be stored as whole units in the lexicon. Similarly, Plank (1976) argues that ample inter- and intra-individual variation makes a rule-based prediction of linking elements in compounds of Germanic languages impossible. This led him to question the existence of generative rules and their cognitive reality.

Japanese also has a partly predictable linking phenomenon. In Japanese compounds, the initial consonant of the second constituent can be voiced, as in */iro/ + /kami/* (color+paper) > */irogami/* (colored paper), a phenomenon referred to as rendaku (Vance, 1980; 1982; 1987). The occurrence of voicing appears not to be predictable (but see below).

Another language with partly-predictable linking elements is Kabardian, a North Caucasian language. In Kabardian, the connectives *-ah-*, *-m-*, *-n-*, and *-r-* can appear between two segments (e.g., p'-ah-ŝ'a = lit. 'nose bottom' 'mustache'). Their use is sometimes facultative and varies dialectally (Kuipers, 1960:76-80).

Let us now turn to linking elements in Dutch noun-noun compounds. The two main linking elements in Dutch are *-s-* and *-en-* (e.g., *schaap+s+kooi* 'sheep fold', *boek+en+kast* 'book shelf'). The linking *-en-* has an orthographic variant, *-e-*. Because both variants are pronounced as schwa in standard Dutch, I will refer to them as *-en-*. The historical origin of *-s-* and *-en-* can be traced back to case endings

to the plural suffix, meaningless.

in medieval Dutch (Booij, 1996; Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997). Synchronically, they are homophonous with the two productive plural markers. Because of this idiosyncrasy, linking elements sometimes carry plural semantics (see below). A statistical survey of the distribution of linking elements in the Dutch part of the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) reveals that of the roughly 23,000 noun-noun compounds which occur at least once in a corpus of 6 million words, 25% contain *-s-*, 11% contain *-en-* and the majority, 65%, occur without any linking element (e.g., *moeder+taal* 'mother tongue'). This distribution is different for compounds in which a derived noun appears as first constituent (17% of all compounds). These compounds almost always occur with a linking element (*-s-*: 62.7%; *-en-*: 32.8%; *-∅-*: 4.6%). In contrast, compounds with first constituents that are not derivations occur mainly without a linking element (71.1%), sometimes with *-s-* (17.3%) and sometimes with *-en-* (11.6%). Even if we leave derivational first constituents aside, linking elements appear too often to be accounted for in terms of exceptions.

One of the reasons that Dutch linking elements are difficult to predict is the possible variation one may observe after one and the same left constituent (e.g., *schaap+en+bout* 'leg of mutton', *schaap+s+kooi* 'sheep fold', and *schaap+herder* 'shepherd'). A relative small set of all first constituents that appear in the CELEX compounds (8.6%) reveal variation in their choice of the linking element. However, they make up 25% of all compounds. In extreme cases, the linking elements vary freely as in the semantically identical words *spelling+verandering* and *spelling+s+verandering* ('spelling change'). First constituents that are derivations show less variation (3.5%) than other first constituents (11.1%). Thus, it is mainly the latter type of first constituents that render prediction problematic.

Given these facts, to which extent can Dutch linking elements be predicted? The linguistic literature lists a range of factors (Van den Toorn, 1981a, 1981b, 1982a, 1982b; Mattens, 1984; Haeseryn et al., 1997; Booij & Van Santen, 1995; Booij, 1996), which can be split into three groups: graded phonological, morphological, and semantic constraints.[2]

An example of a phonological factor is the rule predicting the absence of a linking element after left constituents ending in a vowel, a left constituent ending in a schwa followed by a sonorant, or a left constituent ending in a liquid followed by /k/ or /m/ (*thee+bus* 'tea box', *meubel+zaak* 'furniture shop'). This rule, however,

[2]For a detailed description of all phonological and morphological rules that are proposed in the literature, see Appendix D of chapter 3.

is not without exceptions (*pygmee+en+volk* 'pygmee people'). The use of linking elements is also constrained morphologically, i.e. by a preceding suffix. For instance, the diminutive suffix *-tje* and its variants are always followed by a linking *-s-*. This seems to be the only rule without any counterexamples. Other suffixes reveal apparently unpredictable variation, such as the suffix *-heid* (comparable with English '-ness') which occurs mainly with *-s-*, sometimes without any linker, and in some rare cases with *-en-*. At the semantic level, we can distinguish between rules based on the semantic class of the left constituent and rules based on the semantic relation between the two constituents. If the first constituent is a mass noun, it usually does not occur with a linking element. A counterexample to this rule is *tabak+s+rook* ('tabacco smell'). Similarly, compounds in which the left constituent is the object of a de-verbal agent or action noun to its right also tend to resist insertion of a linking element (*book+verkoper* 'book seller'). Again, this is a rule with exceptions (e.g., *weer+s+verwachting* 'weather forecast').

Ample variation after one and the same constituent and the considerable number of counterexamples to almost any rule that might be proposed has led Van den Toorn (1981a, 1981b, 1982a, 1982b) to conclude that there are no clear rules for Dutch linking elements but only, what he calls, tendencies. The *Algemene Nederlandse Spraakkunst*, the Dutch standard grammar, also speaks of more or less strong tendencies (Haeseryn et al., 1997). Given the literature on Dutch linking elements, one may conclude that their distribution cannot be adequately accounted for by means of traditional linguistic rules.

Previous experimental research on linking elements

Previous experimental studies on linking elements have addressed issues such as the predictability of linking elements, their semantic content, and their morphological status as separate units. This paragraph reviews these studies in chronological order.

The first study is a study by Vance (1980), who addressed rendaku (e.g., */ami/ + /to/* 'net + door' > */amido/* 'screen door') in Japanese novel compounds. Participants had to pronounce compounds in which either the first or the second constituent was a nonword. Vance found a significant correlation between the number of rendaku realizations and the percentage of rendaku in the set of existing compounds that share the right constituent with the target compound, a set I will call the right constituent family. This study thus reveals the first evidence for a paradigmatic effect

of constituent families. We will see that the constituent family, especially the left constituent family, is also the crucial factor for predicting linking elements in Dutch compounds. In addition, I will show that this paradigmatic effect can be computationally formalized and mapped onto a psycholinguistic interactive activation model.

The next experimental study that addressed linking elements in compounds is the paper by Schreuder et al. (1998). This study focuses on the question whether Dutch linking elements have semantic content. Their experiments show that extension of the linking *-e-* to *-en-* interferes with number decision of the whole compound. Thus, the linking *-en-* can activate plural semantics.[3] This study also speaks to the question whether linking elements are processed as separate units. In order to activate plural semantics, linking elements have to be either recognized as separate units, namely as plural suffixes, or combining forms (the left constituent plus linking element) have to be identified as plural forms. However, if plural forms are recognized as whole forms, then their frequency should affect decision latencies. In a post-hoc analysis, Schreuder et al. tested whether there was a correlation between the frequencies of the plural forms and the response latencies. As there was no such correlation, the plural semantics of constituents with *-en-* is probably evoked by the linking element *-en-* that functions as a plural suffix. This indicates that the Dutch linking element *-en-* has a separate representation at the access level.

The next study in time is a study by Kehayia et al. (1999). This study focuses on Polish and Greek noun-noun and adjective-noun compounds that contain linking vowels (e.g., Greek *domat-o-salata* 'tomato salad'; Polish *mebl-o-voz* 'moving van'). They address the questions whether individual constituents are activated during online word recognition, to what extent internal morphological structure plays a role, and whether headedness has an effect on priming. They present a masked priming lexical decision experiment in which compounds were used as target words. Among the different kinds of primes were, for instance, the second constituent, the root of the first constituent, and the combining form (root plus linking vowel). The results show that only existing words can prime a compound. Neither combining forms that are not words nor roots prime. This study does not allow clear conclusions to be drawn about the status of linking elements. More important for this thesis is the finding that left constituents show a significantly stronger priming effect than right constituents. We will see that it is predominantly the left constituent that is crucial

[3]Note that Dressler & Merlini Barbaresi (1994:554-7) propose that, in German, linking elements are semantically empty and have only the function of signaling morphotactic concatenation within a complex word.

for the prediction of linking elements in Dutch compounds.

The most recent study that addresses the issue of linking elements, this time with respect to German compounds, is Dressler, Libben, Stark, Pons, & Jarema (2001). One of their experiments focuses on the question whether the initial constituents, combining forms, or roots are extracted during comprehension. In this experiment, participants had to pronounce the nominative singular form of either the left or the right constituent of a visually presented existing German compound. Response latencies increased with the complexity of the transformation of the presented first constituent into the nominative form. More relevant for the present study is their second experiment, which focuses on the question whether German linking elements are selected by rule or by analogy. For this purpose, participants had to create novel compounds. The authors discuss ten linguistic categories of left constituents that are marked by different gender and final phonemes, and that differ in the choice of linking elements (e.g., schwa-final feminine nouns occur with a linking *-n-* as in *Suppe+n+topf* 'soup+LINK+pot'). The authors propose to determine the appropriate linking element on the basis of eight (in one case six) exemplars for each of the ten categories. In the actual experiment, they selected three left constituents of each category for presentation. Although most of the responses are well predicted by the categories, one category reveals an unexpected number of responses which deviate from the expected linking element. Dressler et al. assume that this variation is due to an analogical effect of the existing compounds that share the first constituent with the target compound, i.e. the left constituent family. As I will show in chapter 6, this analogical effect is stronger than the authors suggest. I will present experiments that show that linking elements in German compounds can be predicted to a considerable degree on the basis of their constituent families.

Aims and outline of the thesis

The main goal of this thesis is to come to a better understanding of how speakers select linking elements in Dutch noun-noun compounds, i.e. linking elements that are only partly predictable by rules. As we have seen, the literature so far lists only tendencies and leaves the prediction problem unsolved. I will address this issue from a psycholinguistic point of view which focuses on the processes and representations involved.

I will propose a formal computational psycholinguistic model of analogy, inspired by the k-NN IB1 algorithm of TiMBL (Daelemans et al., 2000), that captures not

only occasional exceptional and idiosyncratic analogical word formation, but also more systematic morphological patterns and subpatterns. It is a paradigmatic word-type based model for analogy in production, similar in spirit to the model accounting for the also type-based effect of the morphological family in comprehension (Schreuder & Baayen, 1997; De Jong, Schreuder, & Baayen, 2000).

The results of this thesis have interesting implications for the controversy between single-route and dual-route approaches. I will show that, at least in the case of Dutch linking elements, the proposed analogical model has a prediction accuracy that is superior to that of traditional rule-based accounts. On the other hand, my approach shows that it is not necessary to model non-deterministic behavior by means of sub-symbolic representations as in connectionist models. Graded effects can be handled perfectly well with symbolic systems. In addition, since the productive morphological process of selecting a Dutch linking element is better captured by an analogical model, it renders the possibility more plausible that analogy also underlies processes that are traditionally understood as governed by strict syntagmatic symbolic rules.

This thesis is organized as follows. In chapter 2, I will present a first series of production experiments, using the cloze-task, which reveals strong evidence that the choice of linking elements in Dutch novel compounds is analogically determined by the distribution of linking elements in both their left and right constituent families, and that the final suffix of the left constituent (the suffix family) can also affect the choice. Computational simulation studies using TiMBL support the status of the constituent family as the primary basis for analogical prediction. This chapter also presents the outline of a psycholinguistic model for this non-deterministic behavior that does not give up symbolic representations to model non-deterministic form selection.

Chapter 3 presents evidence for another analogical factor, the rime of the left constituent. Production experiments reveal a hierarchical order of left constituent, suffix, and rime. The distribution of the linking elements in the constituent families appears to have the strongest effect on the choice. This effect overrules the suffix and rime effects, while the suffix effect in its turn overrules the rime effect. I model the experimental results with the two exemplar-based computational algorithms, AML and TiMBL. Both models capture the choice of Dutch linking elements in existing and novel compounds with a very similar high degree of prediction accuracy, that compares very favorable with the low prediction accuracy of the rules proposed in the literature.

The factors tested thus far (constituent, suffix, and rime) are all form effects. Chapter 4 focuses on the role of a semantic effect, namely the semantic class of the left and right constituents. I report a production experiment that reveals an effect of animacy and concreteness of the left constituent and no effect whatsoever for the semantic class of the right constituent. Moreover, the form effects of both the left and right constituents appear to be independent of this semantic effect.

Chapter 5 shifts the focus from off-line experiments to reaction time experiments in order to gain insight into the time course of the linker selection. An on-line experiment, in which responses are timed by means of push buttons, reveals the importance of the distribution of the linking elements in the constituent families for the situation in which participants have to respond under time-pressure. This linker decision experiment replicates the effect on the choices of linking elements originally observed using cloze-tasks. Interestingly, it also reveals an analogical effect of the left constituent families on response latencies. In addition, this chapter presents a computational implementation of the psycholinguistic model that was outlined in chapter 2 as an interactive activation network model. A simulation study of the on-line experiment shows that this model captures the effect of the constituent families both on the choices of linking elements as well as on the response latencies.

Chapter 6 addresses the question whether constituent families also affect linking elements in German noun-noun compounds. This study replicates the effect of the left constituent family on the three most common German linking elements, while there seems to be no effect of the right constituent family. Simulation studies with TiMBL show that the selection of German linking elements is affected both by the left constituent family and properties of the left constituent such as rime and gender. As for Dutch, I will outline a psycholinguistic interactive activation model that captures all these factors.

Chapter 7 presents a more general study of Dutch and German compounds. An analysis of the quantitative characteristics of compounds reveals relations between characteristics of words such as frequency and length and the words' occurrence as constituents in complex words. Short and high frequent words appear to be over-represented in complex words.

The findings of chapter 7 provide the statistical tools for the study presented in chapter 8, which focuses on the function of linking elements following derived left constituents, in particular on the function of opening derived left constituents for further word formation, as proposed for German linking elements by Aronoff and Fuhrhop (submitted). I will address this issue by examining the over- and under-

representation of derived nouns as left constituents and the correlation between overrepresentation and frequency as well as productivity in compounds that vary with respect to their linking element.

Finally, chapter 9 summarizes the findings of this thesis and discusses some remaining questions. I will present a comparison of German and Dutch linking elements and I will also examine the possible effect of the stress pattern on Dutch linking elements. Chapter 9 concludes with a discussion of the implications that the findings of this thesis have for the debate on single-route versus dual-route processing.

References

Anshen, F. and Aronoff, M.: 1988, Producing morphologically complex words, *Linguistics* **26**, 641–655.

Aronoff, M. and Fuhrhop, N.: submitted, Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint.

Baayen, R. H. and Renouf, A.: 1996, Chronicling The Times: productive lexical innovations in an English newspaper, *Language* **72**, 69–96.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Booij, G. and Van Santen, A.: 1995, *Morfologie. De Woordstructuur van het Nederlands* (Morphology. The Structure of Dutch Words), Amsterdam University Press, Amsterdam.

Booij, G. E.: 1996, Verbindingsklanken in samenstellingen en de nieuwe spellingregeling (Linking phonemes in compounds and the new spelling system), *Nederlandse Taalkunde* **2**, 126–134.

Botha, R. P.: 1968, *The Function of the Lexicon in Transformational Grammar*, Mouton, The Hague.

Bybee, J. L.: 1985, *Morphology: A study of the Relation between Meaning and Form*, Benjamins, Amsterdam.

Bybee, J. L.: 1988, Morphology as lexical organization, *in* M. Hammond and M. Noonan (eds), *Theoretical Morphology: Approaches in Modern Linguistics*, Academic Press, London, pp. 119–141.

Bybee, J. L.: 1995, Regular morphology and the lexicon, *Language and Cognitive Processes* **10**, 425–455.

Bybee, J. L. and Slobin, D. I.: 1982, Rules and schemas in the development and use of the english past tense, *Language* **58**, 265–289.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 2000, TiMBL: Tilburg Memory Based Learner Reference Guide. Version 3.0, *Technical Report ILK 00-01*, Computational Linguistics Tilburg University.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Dressler, W. U. and Merlini Barbaresi, L.: 1994, *Morphopragmatics. Diminutives and Intensifiers in Italian, German, and Other Languages*, Vol. 76 of *Trends in Linguistics. Studies and Monographs*, Mouton de Gruyter, Berlin, chapter Italian (and German) interfixes, pp. 529–557.

Dressler, W. U., Libben, G., Stark, J., Pons, C. and Jarema, G.: 2001, The processing of interfixed German compounds, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp. 185–220.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Herrera, Z. E.: 1995, *Palabras Estratos y Representaciones: Temas de Fonologia Lexica en Zoque*, El Colegio de Mexico.

Kehayia, E., Jarema, G., Tsapkini, K., Perlak, D., Ralli, A. and Kadzielawa, D.: 1999, The role of morphological structure in the processing of compounds: The interface between linguistics and psycholinguistics, *Brain and Language* **68**, 370–377.

Kuipers, A. H.: 1960, *Phoneme and Morpheme in Kabardian*, Mouton and Co., The Hague.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R. and Pinker, S.: 1995, German inflection: The exception that proves the rule, *Cognitive Psychology* **29**, 189–256.

Mattens, W. H. M.: 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343.

Pinker, S.: 1991, Rules of language, *Science* **153**, 530–535.

Pinker, S.: 1997, Words and rules in the human brain, *Nature* **387**, 547–548.

Pinker, S.: 1999, *Words and Rules: The Ingredients of Language*, Weidenfeld and Nicolson, London.

Pinker, S. and Prince, A.: 1991, Regular and irregular morphology and the psychological status of rules of grammar, *Proceedings of the 1991 meeting of the Berkeley Linguistics Society*.

Plank, F.: 1976, Morphological aspects of nominal compounding in German and certain other languages: what to acquire in language acquisition in case the rules fail?, *in* G. Drachman (ed.), *Akten des 1. Salzburger Kolloquiums über Kindersprache*, number 2 in *Salzburger Beiträge zur Linguistik*, Gunter Narr,

Tübingen, pp. 201–219.

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S. and Mars, F.: 1997, Morphological priming, fragment completion, and connectionist networks, *Journal of Memory and Language* **36**(3), 382–405.

Rumelhart, D. E. and McClelland, J. L. (eds): 1986, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, MIT Press, Cambridge, Mass.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Schreuder, R., Neijt, A., Van der Weide, F. and Baayen, R. H.: 1998, Regular plurals in Dutch compounds: linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Skousen, R.: 1989, *Analogical Modeling of Language*, Kluwer, Dordrecht.

Skousen, R.: 2000, Analogical modeling and quantum computing, Los Alamos National Laboratory <http://arXiv.org>.

Unbegaun, B.: 1967, *Russian Grammar*, Clarendon Press.

Van den Toorn, M. C.: 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is (Linking phonemes in compounds with derived forms as first constituents), *De Nieuwe Taalgids* **74**, 197–205.

Van den Toorn, M. C.: 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is (Linking phonemes in compounds with loanwords as first constituents), *De Nieuwe Taalgids* **74**, 547–552.

Van den Toorn, M. C.: 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I), *De Nieuwe Taalgids* **75**(1), 24–33.

Van den Toorn, M. C.: 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for the regulation of linking phonemes in nominal compounds II), *De Nieuwe Taalgids* **75**(2), 153–160.

Vance, T. J.: 1980, The psychological status of a constraint on Japanese consonant alternations, *Linguistics* **18**, 145–167.

Vance, T. J.: 1982, On the origin of voicing alternation in Japanese consonants, *Journal of the American Society* **102**(2), 333–341.

Vance, T. J.: 1987, *An Introduction to Japanese Phonology*, State University of New York, Albany, chapter Sequential voicing, pp. 13–148.

Constituent families

This chapter has been published as Andrea Krott, Robert Schreuder, and R. Harald Baayen· 2001, Analogy in morphology: modeling the choice of linking morphemes in Dutch, *Linguistics* 39 (1), 51-93.

Abstract

This study argues that a productive, but not fully-regular morphological phenomenon, the choice of linking morphemes in Dutch nominal compounds, is based on analogy. In Dutch, a linking *-s-* or *-en-* can appear between the constituents of a nominal compound. We present production experiments which reveal strong evidence that the choice of linking morphemes in novel compounds is analogically determined by the distribution of linking morphemes in what we call the 'constituent families'. A 'constituent family' is the set of existing compounds that share the first (or second) constituent with the novel compound. A further experiment shows that in the case of derived pseudo-words as first constituents, it is the family of the suffix which influences the choice of the following linking morpheme. In addition to these experiments, we present computational simulation studies in which the choices made by participants in our experiments are predicted with a high degree of accuracy using a machine-learning algorithm for analogy. These studies support the status of the constituent family as the primary basis for analogical prediction. Finally, we outline a psycholinguistic model for analogy in the mental lexicon that does not give up symbolic representations and, at the same time, captures non-deterministic variation.

Introduction

Morphological variation can often be captured by simple rules. Consider, for example, the realization of the regular plural of English nouns, which appears in three different forms, /ɪz/, /z/, and /s/. These three variants can be predicted on the basis of the phonological form of the base word. The plural is pronounced /ɪz/ after bases ending in sibilants (e.g., *horses*), /z/ after bases ending in vowels and voiced segments other than /z/, /ʒ/, and /dʒ/ (e.g., *beds*), and it is pronounced /s/ after bases ending in voiceless segments other than /s/, /ʃ/, and /tʃ/ (e.g., *months*).

In addition to this kind of regular variation, there are morphological domains where the choice between alternative realizations is less predictable. One such domain is the analysis of linking elements in compounds, which are also referred to as connectives, interfixes, linkers, or linking morphemes. Linking elements occur in various languages across different language families. In English, linking elements are extremely rare. We know of only a few examples, all built with the head word *man*: *marksman*, *sportsman*, *craftsman*, *kinsman*, *tradesman*, and *spokesman*. The last example, in which the *-s-* appears without any possible semantic function, best illustrates the phenomenon of linking elements. In some languages, linking elements can be fully predicted on the basis of the phonological characteristics of the preceding (and/or the following) constituent. For instance, Zoque, a Mixe-Zoquean language spoken in Mexico, has a nominal compound formation in which the linking element is a vowel that is identical to the vowel in the preceding syllable. However, in many other languages such clear rules cannot be formulated. For example, Kabardian (North Caucasian) has the linking elements *-ah-*, *-m-*, *-n-*, and *-r-*, which tend to be obligatory in some morphological contexts and optional in others (Kuipers, 1960:78–80). In Indo-European, the Germanic languages are especially rich in non-predictable or only partly predictable variation in the use of linking elements (e.g., Danish, Norwegian, Swedish, and German). The distribution of the two main linking elements in Dutch, *-en-* and *-s-*, is likewise only partially predictable by rule.

The systematicities governing the selection of linking morphemes is a longstanding unsolved problem in the morphology of Dutch and many other Germanic and Non-Germanic languages. It is an issue that has hardly received attention in the generative tradition,[1] with the exception of Botha (1968), even though it is a problem

[1] For instance, the 800 page handbook of morphology edited by Spencer & Zwicky (1998) devotes five lines of text to the problem of linking elements (p.81).

that receives discussion in any good reference grammar (e g , Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997, de Haas & Trommelen, 1993)

A first goal of the present study is to show that the distribution of linking morphemes in Dutch noun-noun compounds can be accounted for by means of a formal computational model of analogy with a higher degree of observational adequacy than can be achieved by means of the rules proposed in the literature Our conclusions are based on both surveys of existing compounds in the Dutch lexicon as well as on the choices for linking morphemes in novel compounds as produced by participants under strict experimental conditions

A second goal is to contribute to the discussion in the current literature about the nature of morphological rules, whether such rules are symbolic in nature (Clahsen, 1999, Marcus, Brinkman, Clahsen, Wiese, & Pinker, 1995, Pinker, 1991, 1997) or whether rules are an epiphenomenon of distributed storage in connectionist networks (Seidenberg, 1987, Seidenberg & Hoeffner, 1998, Plunkett & Juola, 1999, Rueckl, Mikolinski, Raveh, Miner & Mars, 1997) The phenomenon that we are dealing with is interesting in the sense that it is fully productive and yet not completely regular As such, it poses a serious challenge to proponents of symbolic rule systems At the same time, we will show that it is possible to predict non-deterministic aspects of human cognition without necessarily making use of distributed connectionist networks In this sense, our present analogy-based approach provides an alternative to both symbolist and connectionist approaches to cognition

The notion of analogy as we use it in this paper is different from its two traditional interpretations in linguistics First, analogy is often contrasted with rules, with regular novel forms being formed by rules, and exceptional novel forms being built by analogy to individual examples (e g , *brunch* by analogy to *smog*, see, e g , Anshen & Aronoff, 1988) Second, analogy can also be understood as the initial basis for the acquisition of rules In this view, analogical learning might be involved in determining the conditions under which a rule applies But once a rule is established, the instances which led to the rule would then be irrelevant, and would not be kept in memory

Our use of the term analogy differs from these two interpretations in the following ways First, the kind of analogy with which we are concerned is not the kind of analogy that occasionally leads to exceptional new creatively coined words such as *brunch* Instead, we are concerned with the regular phenomena that are traditionally described by means of linguistic rules Following Skousen (1989) and Daelemans, Zavrel, Van der Sloot, & Van den Bosch (1999), we adopt a formal and

computationally tractable definition of analogy that offers a new way of understanding the way in which linguistic rules actually work. Second, we hypothesize that, at least in the domain of morphological processing, there are no rules that are formed on the basis of initially stored examples of complex words, with the initial exemplars fading from memory. Instead, we assume that many fully regular complex words, both inflected and derived, remain available in the mental lexicon (e.g., Bertram, Laine, Baayen, & Schreuder, 1999; Bertram, Schreuder, & Baayen, 2000; Baayen, Dijkstra, & Schreuder, 1997; Sereno & Jongman, 1997; Sandra, Frisson, & Daems, 1999; Taft, 1979; Baayen, Schreuder, De Jong, & Krott, in press), and serve as exemplars for the analogical formation of novel forms. In other words, we hypothesize that rules are essentially analogical in nature (De Saussure, 1966).

In what follows, we first describe the problem of the systematicities underlying the distribution of linking morphemes in Dutch, and we show that the notion of default rules that has figured prominently in recent discussions (Marcus et al., 1995; Clahsen, 1999) is not applicable to this phenomenon. In the next section, we present the results of three production experiments, which show that, the substantial variation in the choice of linking morphemes notwithstanding, Dutch native speakers tend to converge on the same linking elements for novel compounds. These experiments show, furthermore, that the choice of a linking element for a novel compound is strongly influenced by the distribution of linking elements in the set of existing compounds sharing the first or second constituent with the novel compound (e.g., *fiets* 'bike' in *fiets-pad* 'cycle path' and *fiets+bel* 'bicycle bell', and *winkel* 'shop' in *schoen+winkel* 'shoe shop' and *hoed+en+winkel*, 'hat+PLUR+shop', 'hat shop'). We will refer to these sets of compounds as constituent families.

In the subsequent section, we will show that the notion of analogy based on constituent families can be formalized computationally, and that this allows us to predict the distribution of linking morphemes in the Dutch lexicon and also to predict the performance of our experimental participants. In the general discussion, we outline how the computational model can be mapped onto a psycholinguistically more realistic spreading activation model along the lines of Schreuder & Baayen (1995).

Linking morphemes in Dutch: no rules but tendencies

In this section, we describe the distributional properties of the linking elements in Dutch and their linguistic status. The two main linking elements in Dutch noun-noun compounds are *-s-* and *-en-*. The latter is occasionally realized in the orthography as *-e-*. Both *-en-* and *-e-* are pronounced as schwa in standard Dutch. As the present study focuses on the production of linking elements, we do not distinguish between the two orthographic realizations.

There is a long-standing discussion about the status of these linking elements. Are they just meaningless letters or do they carry semantic information? Both *-s-* and *-en-* are homographic with the two productive plural suffixes of Dutch nouns.[2] The linking element *-en-* may only appear after left constituents that themselves pluralize with *-en*. The linking element *-s-* is not constrained in the same way. It may appear following constituents with which it does not form a plural. There is evidence that *-en-* marks plurality in compounds, as shown by Schreuder, Neijt, Van der Weide, & Baayen (1998). Neijt, Baayen, & Schreuder (in preparation) show that, depending on the first constituent, the *-s-* may also convey plural semantics. In the light of this evidence, we will henceforth refer to *-en-* and *-s-* as linking morphemes rather than linking elements. Note, however, that the question whether the *-s-* and *-en-* forms in Dutch compounds are indeed completely identical to the Dutch plural suffixes is not what is at issue in the present study. Our aim here is to come to grips with the distribution of these forms irrespective of their morphological status.

The literature on linking morphemes in Dutch compounds has attempted to capture the distribution of linking morphemes by means of rules operating at the levels of phonology, morphology, and semantics (see, e.g., Van den Toorn, 1981a; 1981b; 1982a; 1982b; Mattens, 1984). An example of a phonological rule is the constraint that after first constituents ending in a vowel, or ending in a schwa followed by a sonorant, or ending in a liquid followed by /k/ or /m/ (*thee+bus* 'tea box'; *meubel+zaak* 'furniture shop'), linking morphemes are not allowed. This rule is not without exceptions, however, as shown by a compound such as *pygmee+en+volk*, 'pygmy+PLUR+people', 'pygmy people'.

At the morphological level, particular affixes show preferences for specific linking

[2]Marcus et al (1995) and Clahsen, Eisenbeiss & Sonnenstuhl-Henning (1997) have argued that it is impossible for a language to have more than one productive rule for a particular inflectional function. This claim is based on the distribution of noun plurals in German. The Dutch plural system provides a counterexample to this claim, as shown by Baayen, Schreuder, De Jong, & Krott (in press), a study that presents detailed linguistic and psycholinguistic evidence for the regularity and productivity of both Dutch plural suffixes.

morphemes For instance, the diminutive suffix *-je* is always followed by the linking *-s-* in nominal compounds (*plaat+je+s+boek*, picture+DIMINUTIVE+PLUR+book 'small pictures book') Other morphemes show strong preferences, such as the suffix *-heid*, '-ness', which appears predominantly with *-s-*, but occasionally without a linking morpheme and rarely with *-en-*

At the level of semantics two different kinds of constraints have been observed First, the semantics of the first constituent may render the use of a linking morpheme unlikely For instance, mass nouns are not followed by linking morphemes (e g , *papier+handel* 'paper trade', exception *tabak+s+rook*, 'tabacco+GENITIVE+ smoke', 'tabacco smoke') Conversely, the linking morpheme *-en-* often occurs when the first constituent of a compound has a plural interpretation (Haeseryn et al , 1997 685, Schreuder et al , 1998) *boek+en+kast*, 'book+PLUR+case', 'book case', *krent+en+brood*, 'currant+PLUR+bread', 'currant bread', exception *boek+ handel*, 'book shop' Semantic factors may interact with the morphological structure of the first constituent For instance, first constituents ending in *-er* denote human agents or objects For human agents one tends to find the linking *-s-*, as in *duik+er+s+ziekte*, 'dive+er+PLUR+sickness', 'decompression sickness', while for inanimate objects one tends to find no linking morpheme, as in *straal+jager+piloot*, 'stream+hunt+er+pilot', 'fighter jet pilot' These rules are also not without exceptions (e g , *leraar+en+opleiding* ('teacher+PLUR+education', 'education of teachers')) (see Mattens 1984) Second, the semantic relation between the two constituents has also been argued to codetermine the choice of the linking morpheme For instance, copulative compounds such as *man+wijf*, 'man+bitch', 'mannish woman' never take a linking morpheme Similarly, compounds in which the first constituent is the object of a de-verbal agent or action noun to its right also tend to resist insertion of linking morphemes (*boek+verkoper* 'book seller', exception *weer+s+ verwachting*, 'weather+GEN +expectation', 'weather forecast')

A final property of linking morphemes in Dutch is that they evidence a certain amount of variability For instance, the word 'spelling change' has two translation equivalents in Dutch, *spelling+verandering* and *spelling+s+verandering* Even for a single speaker, forms such as these appear to be in free variation

Summing up, first constituents seem to have the strongest influence on the choice of linking morphemes, phonologically, morphologically, and semantically The second constituent plays a minor role, being a codeterminant of the semantic relation between the two constituents The numbers of exceptions to the rules describing the distribution of linking morphemes are so large that Van den Toorn (1982) has

argued that we are dealing with tendencies rather than with real rules.

It is important to note that the distribution of the linking morphemes in Dutch does not lend itself to an analysis in terms of a set of rules including a default rule. In such a system of rules, a series of positively specified cases is supplemented by a general case, the default, for which a simple and straightforward definition of its input domain (in the sense of Van Marle, 1985) cannot be given.[3] Focussing on the phonological rules for the distribution of Dutch linking morphemes, we observe only negative specifications: Linking morphemes do not appear following left constituents that end in a vowel, in a schwa followed by a sonorant, or in a liquid followed by /k/ or /m/. Crucially, the notion of a default, covering those words that do not fall under the negatively specified input domains, does not make sense for Dutch linking morphemes, as it does not have any predictive power with respect to the appropriate linking morpheme. Thus, words falling under the default, i.e. words that do not end in a vowel, in a schwa followed by a sonorant, or in a liquid followed by /k/ or /m/, can still appear in a compound with no linking morpheme, with *-s-* or with *-en-*. Clearly, none of these three possibilities can be the default choice. Turning to the level of morphology, we again find that the notion of a default is not applicable, as each suffix has its own stronger or weaker preferences. Similarly, at the level of semantics, we only observe random subgeneralizations without a well specified overall default. In spite of the absence of a rule system with a default, speakers of Dutch nevertheless have strong intuitions about which linking morpheme is appropriate for novel compounds.

Production experiments

In this section, we address two related questions. First, to what extent do native speakers of Dutch agree about which linking morphemes are most appropriate to use in novel compounds? How much variability can be observed given the strong intuitions of native speakers as to what might be the appropriate choice? Second, what factors underlie these strong intuitions? We shall see that there is indeed strong agreement about which linking morpheme is most appropriate. As to the factors underlying the choice of linking morphemes, we shall see that the existing compounds sharing the left (or right) constituent with the target compound form perhaps the most important factor of all. In what follows, we will refer to these compounds as the left and right constituent families of such a target compound.

[3]For an analysis of German noun pluralization in such a framework, see Marcus et al., 1995.

An individual compound in such a family will be referred to as a constituent family member

The next section presents experimental evidence for the important role of the constituent families for the linking morphemes *-en-* and *-s-* The following section investigates the relevance of the morphological structure of the first constituent We have not explicitly included semantic and phonological factors in our experimental design However, we will show that analogical modeling of the experimental data yields slightly better results when semantic properties of the constituents are also taken into account Including phonological information results in slightly worse performance

The constituent family effect

The next two subsections present experiments studying the effect of the constituent family on the choice of the linking morphemes *-en-* and *-s-*

Experiment 1: The linking morpheme -en-

If the choice of linking morphemes in novel compounds were based simply on the distribution of the linking morphemes in the lexicon as a whole, one would expect speakers to choose not to use a linking morpeme in roughly 7 out of 10 cases 69% of all compounds listed in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) appear without any linking morpheme Their second best guess would then be *-s-*, which occurs in 20% of the compounds in this database, and their least probable bet would be *-e(n)-* (11%) In the light of the linguistic description of the distribution of *-en-* and *-s-* presented in the previous section, this simple guessing behavior is unlikely On the other hand, the linguistic rules that have been formulated tend to have so many exceptions that their explanatory value is called into question as well In what follows, we explore the hypothesis that native speakers of Dutch base their choice on the relative frequencies of the linking morphemes as realized not in the lexicon as a whole, but in the restricted sets comprising the constituent families of individual compounds

Method

Materials We constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3) Each set contained 21 nouns The constituents of L1 and R1 had constituent families with as strong a bias as possible towards the

linking morpheme *-en-*. Conversely, L3 and R3 showed a bias as strong as possible against *-en-*, though we made sure that these constituents form their plural with the suffix *-en* so that a linking *-en-* is possible. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against *-en-*. We used the CELEX lexical database (Baayen et al., 1995) to determine the constituent families of the constituents in these six sets. Compounds with a token frequency of zero in a corpus of 42 million words were not included.

The constituents in the L1 set had constituent family members of which at least 70% contained the linking morpheme *-en-*. The mean number of compounds in these families was 12.5 (range 5–43). Their mean token frequency was 149.2 per 42 million wordforms (range 58–439). The range of choices for R1 constituents was more restricted. The constituents in the R1 set therefore had constituent family members of which at least 60% contained the linking morpheme *-en-*. The mean number of compounds in these families was 3.6 (range 2–7). Their mean token frequency was 49.1 per 42 million wordforms (range 20–119). Neutral left constituents are rare. The neutral set L2 included left constituents whose families contained between 35% and 65% compounds with the linking morpheme *-en-*. These families had a mean number of compounds of 8.3 (range 3–24) and a mean token frequency of 136.3 per 42 million wordforms (range 15–439). The constituents in the R2 set had constituent family members of which 40% to 60% contained the linking morpheme *-en-*. These families had a mean number of compounds of 5.3 (range 3–15) and a mean token frequency of 66.7 per 42 million wordforms (range 8–192). The remaining sets L3 and R3, the groups with a bias against *-en-*, contained constituents whose family members never have a linking *-en-*. There were in the mean 25 (range 11–66; L3) and 17.9 (range 10–47; R3) family members respectively. Their mean token frequency was 573.7 (range 98–2650; L3) and 349.8 (range 47–2290; R3). These are the maximal contrasts that allowed us to select 21 constituents for each experimental set.

Each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: Bias in the Left Position (Positive, Neutral, and Negative) and Bias in the Right Position (Positive, Neutral, and Negative). None of these compounds is attested in the CELEX lexical database with a token frequency higher than zero. All have a high degree of semantic interpretability. Appendix A lists all experimental items. The 9 x 21 = 189 experimental items were divided over three lists. List 1 contained the compounds of the facto-

rial combinations L1-R1, L2-R3, and L3-R2. List 2 contained the compounds of the combinations L1-R2, L2-R1 and L3-R3, and List 3 contained the compounds of the combinations L1-R3, L2-R2, and L3-R1. In this way, each participant saw a given constituent only once. We constructed a separate randomized list of the 3 x 21 = 63 compound constituent pairs for each participant.

Procedure. The participants performed a cloze-task. The experimental list of items was presented to the participants in written form. Each line presented two compound constituents separated by two underscores. We asked the participants to combine these constituents into new compounds and to specify the most appropriate linking morpheme, if any, at the position of the underscores, using their first intuitions. Occasionally, the first constituent may change its form when it is combined with a linking morpheme (e.g., *schip* ('ship') appears as *scheep* in the compound *scheepswerf* ('shipyard')). The instructions made clear that these changes were not of interest and could be ignored. We told the participants that they were free to use *-en-* or *-e-* as spelling variants of the linking morpheme *-en-*. The experiment lasted approximately 15 minutes.

Participants. Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch. The participants were divided into three groups. Each group was asked to complete one of the three experimental lists.

Results and discussion

Occasionally, participants filled in a question mark or a letter sequence other than a linking morpheme. Such responses were counted as errors. The overall error rate was extremely low (0.05%), which allowed us to include all participants and all items in the data analysis. Table 2.1 summarizes the percentages of *en* responses versus other responses for the nine experimental conditions. Appendix A lists the individual words together with the absolute numbers of *en* and *not en* responses.

A by-item logit analysis (see, e.g., Rietveld & Van Hout, 1993; Fienberg, 1980) of the *en* and *not en* responses revealed a main effect of Bias in the Left Position ($F(2,180) = 119.3$, $p < .0001$), a main effect of Bias in the Right Position ($F(2,180) = 12.8$, $p < .0001$), and no interaction of the Bias in both positions ($F(4,180) < 1$). Although the Neutral Bias condition for the right constituents led to slightly higher numbers of *en* responses than the Positive Bias condition, the difference between these two conditions is not reliable ($F(1,120) = 1.1$, $p = .2974$).

The upper panel of Figure 2.1 shows the effects of both biases on the percentage

Table 2.1: Percentages of selected linking morphemes when varying bias for *-en-* (Positive, Neutral, and Negative) in the left and right compound position. Standard deviations between parentheses.

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	en	94.8	(11.2)	96.4	(6.7)	87.4	(15.3)
	not en	5.2	(11.2)	3.6	(6.7)	12.6	(15.3)
Neutral							
	en	75.0	(23.7)	81.9	(15.5)	58.3	(26.9)
	not en	25.0	(23.7)	18.1	(15.5)	41.2	(26.9)
Negative							
	en	18.1	(19.1)	18.8	(19.9)	6.0	(7.7)
	not en	81.9	(19.1)	81.2	(19.9)	94.0	(7.7)

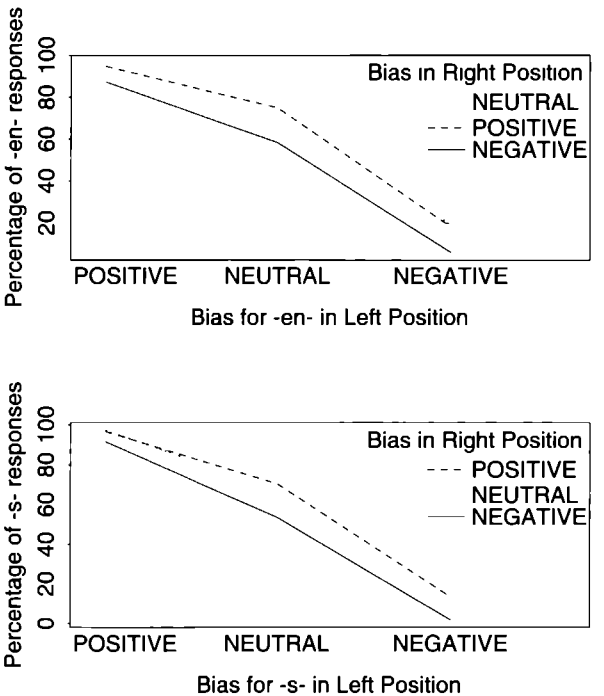


Figure 2.1: Interaction of Biases in Left and Right Position for the linking morphemes *-en-* (upper panel) and *-s-* (lower panel).

of *en* responses Bias has a larger effect on the Left Position (a difference of roughly 80% between the Positive and Negative conditions) than on the Right Position (a difference of roughly 15%) This result reflects an asymmetry in the distribution of the linking elements in Dutch that is also mirrored in our experimental design

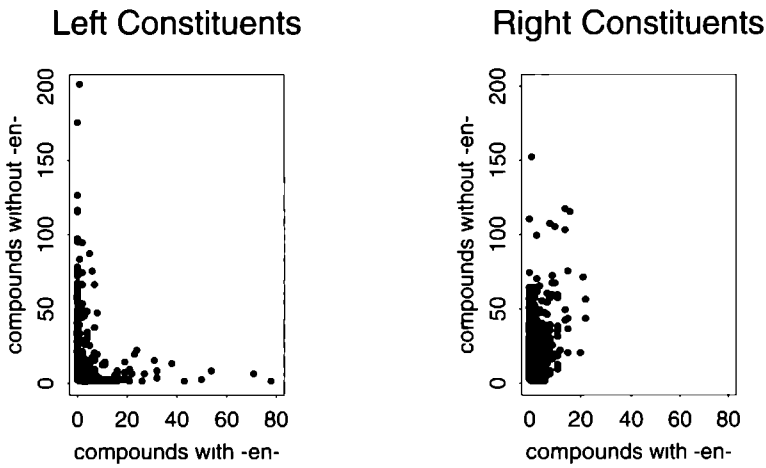


Figure 2 2 Distribution of numbers of compounds with and without the linking morpheme *-en-* for left and right constituents

Figure 2 2 illustrates this asymmetry for the families of left and right constituents of compounds with the linking morpheme *-en-* The left panel is a scattergram for the left constituents It represents each of the 4320 constituents by a dot in the plane spanned by the number of compounds with *-en-* in which it appears (horizontal axis) and the number of compounds without *-en-* in which it appears (vertical axis) Note that the points are scattered along the two axes, indicating that there are many left constituents that occur predominantly either with *-en-* or without *-en-* Turning to the right panel, we find a more random pattern for the 3935 right constituents Here, the presence of a larger number of compounds with *-en-* does not imply a small number of compounds without *-en-*, and vice versa Thus, a strong bias for *-en-* exists only for left constituents Interestingly, this asymmetry is clearly reflected in the responses of the participants of the present experiment If participants had chosen the linking morpheme at random on the basis of all the existing compounds (CELEX 43413) in the language, one would have expected *-en-* (CELEX 4744) to be selected in roughly 11% of our experimental material The

left constituents provide larger families with clearer preferences for or against *-en-*, leading to a much higher percentage of *en* responses in the Positive and Neutral conditions (58%–96% versus 6%–19% in the Negative condition).

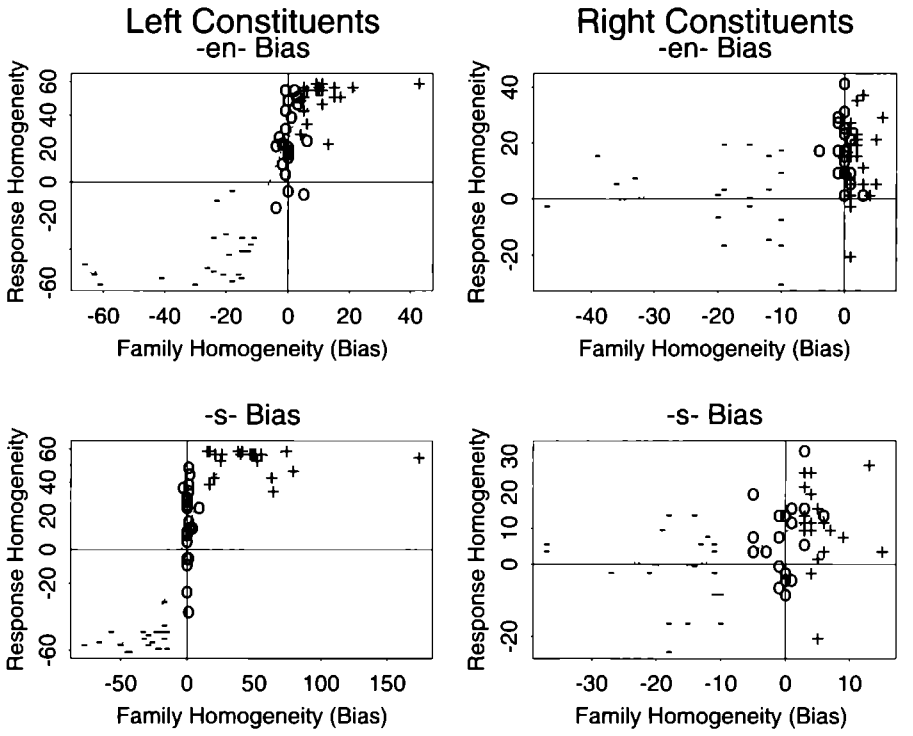


Figure 2.3: Correlation of *-en-* and *-s-* family homogeneity and *-en-* and *-s-* response homogeneity with local smooth lines; '+': Positive Bias, 'o': Neutral Bias, '-': Negative Bias.

In a post-hoc analysis we also tested the overall effect of family homogeneity on the response homogeneity across the three conditions (Positive, Neutral, Negative) both for the Left and Right Bias. We calculated the family homogeneity in terms of the difference between the number of family members with *-en-* and the number of family members without *-en-*. We calculated the response homogeneity in terms of the difference between the number of *en* responses and other responses.

The upper panels of Figure 2.3 reveal a non-linear correlation between response homogeneity and family homogeneity represented by a dotted line.[4] The upper left

[4]We used a non-parametric regression smoother (see Cleveland, 1979), as parametric techniques based on linear models are clearly inappropriate for our data.

panel shows a sigmoid curve for the left constituents. The upper right panel shows a more diffuse pattern for the right constituents. Despite this difference, a Spearman correlation test revealed a significant correlation between the family homogeneity and the response homogeneity both for the Left ($r_s = .87$, $z = 6.88$, $p < .0001$) and the Right Position ($r_s = .34$, $z = 2.70$, $p = .007$). The magnitude of these correlation coefficients ($r_s = .87$ versus $r_s = .34$) correspond to the difference in strength of the Left and Right Bias: In terms of rank correlations, the Left Bias explains 76% of the variance, while the Right Bias explains only 12% of the variance.

Having observed clear effects of analogy on the choice of the linking morpheme *-en-*, we now turn to the linking morpheme *-s-*.

Experiment 2: The linking morpheme -s-

Method

Materials. As in Experiment 1 we constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3). Each set contained 21 nouns. The constituents of L1 and R1 sets had constituent families with as strong a bias as possible towards the linking morpheme *-s-*. Conversely, L3 and R3 showed a bias as strong as possible against *-s-*. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against *-s-*. We used the CELEX lexical database to determine the constituent families of the constituents in these six sets. Compounds with a token frequency of zero in a corpus of 42 million words were not included.

The constituents in the L1 set had constituent family members of which at least 80% contained the linking morpheme *-s-*. The mean number of compounds in these families was 45.7 (range 15–174). Their mean token frequency was 1196.8 per 42 million wordforms (range 102–6663). The constituents in the R1 set had constituent family members of which at least 70% contained the linking morpheme *-s-*. The mean number of compounds in these families was 6.5 (range 4–19). Their mean token frequency was 103.5 per 42 million wordforms (range 12–409). Neutral left constituents are rare. The neutral set L2 included left constituents whose families contained between 35% and 65% compounds with the linking morpheme *-s-*. These families had a mean number of compounds of 6.4 (range 2–34) and a mean token frequency of 116.9 per 42 million wordforms (range 5–915). The constituents in the R2 set had constituent family members of which 45% to 55% contained the linking morpheme *-s-*. These families had a mean number of compounds of 16.4 (range 4–52) and a mean token frequency of 216.4 per 42 million wordforms (range

18–527). The remaining sets L3 and R3, the groups with a bias against *-s-*, contained constituents whose family members never have a linking *-s-*. There were in the mean 31.2 (range 15–77; L3) and 2.45 (range 10–37; R3) family members respectively. Their mean token frequency was 903.1 (range 98–2874; L3) and 532.9 (range 39–2677; R3). These are the maximal contrasts that allowed us to select 21 constituents for each experimental set.

As in Experiment 1, each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: Bias in the Left Position (Positive, Neutral, and Negative) and Bias in the Right Position (Positive, Neutral, and Negative). None of these compounds is attested in the CELEX lexical database with a token frequency higher than zero. All have a high degree of semantic interpretability. Appendix B lists all experimental items. The 9 x 21 = 189 experimental items were divided over three lists. List 1 contained the compounds of the factorial combinations L1-R1, L2-R3, and L3-R2. List 2 contained the compounds of the combinations L1-R2, L2-R1 and L3-R3, and List 3 contained the compounds of the combinations L1-R3, L2-R2, and L3-R1. In this way, each participant saw each constituent only once. We constructed a separate randomized list of the 3 x 21 = 63 compound constituent pairs for each participant.

Procedure. The procedure was identical to that of Experiment 1.

Participants. Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch, none had participated in the previous experiment. The participants were divided into three groups. Each group was asked to complete one of the three experimental lists.

Results and discussion

The participants followed the instructions very closely so that no responses had to be counted as errors. That allowed us to include all participants and all items in the data analysis. Table 2.2 summarizes the percentages of *s* responses versus other responses for the nine experimental conditions. Appendix B lists the individual words together with the absolute numbers of *s* and *not s* responses.

A by-item logit analysis of the *s* and *not s* responses revealed a main effect of Bias in the Left Position ($F(2,180) = 150.6$, $p < .0001$), a main effect of Bias in the Right Position ($F(2,180) = 10.5$, $p < .0001$), and no interaction of the Bias in both positions ($F(4,180) = 1.6$, $p = .1883$). Again, the difference between the Neutral

Table 2.2: Percentages of selected linking morphemes when varying bias for *-s-* (Positive, Neutral, and Negative) in the left and right compound position. Standard deviations between parentheses.

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	s	96.7	(20.3)	97.4	(22.8)	91.7	(24.2)
	not s	3.3	(20.3)	2.6	(22.8)	8.3	(24.2)
Neutral							
	s	70.5	(10.2)	67.6	(3.7)	53.6	(9.5)
	not s	29.5	(10.2)	32.4	(3.7)	46.4	(9.5)
Negative							
	s	13.6	(5.2)	5.2	(11.5)	1.9	(5.1)
	not s	86.4	(5.2)	94.8	(11.5)	98.1	(5.1)

and Positive Bias conditions on the Right Position is not reliable ($F(1,120) = 1.9$, $p = .1687$).

The lower panel of Figure 2.1 shows the effects of both Biases on the percentage of *s* responses. As in Experiment 1, Bias has a larger effect on the Left Position (a difference of minimal 70% between the Positive and Negative conditions) than on the Right Position (a difference of maximal 17%). This result again reflects an asymmetry in the distribution of the linking elements in Dutch that is also mirrored in our experimental design. The left constituents provide larger families with clearer preferences for or against *-s-*, leading to a much higher percentage of *s* responses in the Positive and Neutral conditions (from 53% up to 97% versus 2% up to 14% for the Negative condition).

In a post-hoc analysis we tested the overall effect of the family homogeneity on the response homogeneity across the three conditions (Positive, Neutral, Negative) both for the Left and Right Bias. As before, we calculated the family homogeneity in terms of the difference between the number of family members with *-s-* and the number of family members without *-s-*. We calculated the response homogeneity in terms of the difference between the number of *s* responses and other responses. The lower panels of Figure 2.3 reveal a non-linear correlation between response homogeneity and family homogeneity represented by a dotted line. The lower left panel shows the data of the left constituents, the lower right panel shows the data of the right constituents. As for the *-en-* homogeneity, the left constituents

reveal a sigmoid curve, while the right constituents show a more diffuse pattern. As in Experiment 1, a Spearman correlation test revealed a significant correlation between the family homogeneity and the response homogeneity both for the Left (r_s = .89, z = 7.00, $p < .0001$) and the Right Position (Spearman: r_s = .42, z = 3.33, $p < .0001$). The magnitude of these correlation coefficients (r_s = .89 versus r_s = .42) correspond to the difference in strength of the Left and Right Bias: In terms of rank correlations, the Left Bias explains 79% of the variance, while the Right Bias explains only 18% of the variance.

Experiment 2 addressed the question whether the families of the right and left constituent affect the choice for or against the linking morpheme *-s-* when building a new nominal compound. We were able to replicate the results of Experiment 1 which tested the family effect on the linking morpheme *-en-*. The family of the left constituent has a strong effect on the choice of the linking morpheme, while the family of the right constituent has a smaller, but also significant effect.

The suffix family effect

Experiment 3: The effect of the preceding suffix on the linking -s-

We have seen that the families of the immediate constituents of a new nominal compound have a great influence on the choice of the linking morpheme. The linguistic literature tells us that, in the case of derived words as left constituents, it is the suffix that has influence on the following linking morpheme (Van den Toorn, 1981a; 1981b). For instance, suffixes *-ist* (similar to English person-noun forming '-ist') or *-in* (similar to English '-ess') appear mainly with *-en-*, while suffixes *-aard* (similar to English '-ee') or *-heid* (similar to English '-ness') appear mainly with *-s-*. However, like the constituents, the suffix does not completely determine the linking morpheme. We therefore tested whether the suffix family, i.e. all compounds which contain a left constituent built with a particular suffix, has an effect on the choice of the linking morpheme. For this experiment we chose the linking morpheme *-s-* because the *-s-* appears much more often with a preceding suffix (586/1004 = 58.4% of all preceding derived words) than the *-en-* (54/594 = 9.1% of all preceding derived words). To make sure that we test the effect of the suffix and not the effect of the left constituent, we used pseudo-derivations.

Method

Materials. We constructed two sets of left pseudo-constituents (L1, L2) and three sets of right existing constituents (R1, R2, R3). Each set contained 21 nouns. The pseudo-constituents of the sets L1 and L2 contained Dutch suffixes with pseudo-stems, none of which violated the phonotactic rules of Dutch. The suffixes of L1 were *-ing* (similar to English '-ing'), *-heid* (similar to English '-ness'), and *-iteit* (similar to English '-ity'). They appear in CELEX compounds mainly with the linking morpheme *-s-* (*-ing*: 379/406 = 93.3%; *-heid*: 65/66 = 98.5%; *-iteit*: 21/25 = 84.0%). The suffixes of L2 were *-in* (similar to English '-ess'), *-sel* (similar to English '-ee'), and *-ster* (similar to English '-ess'). They appear in CELEX in at least 50% without the linking morpheme *-s-* (*-in*: 0/1 = 0%; *-sel*: 0/6 = 0%; *-ster*: 1/2 = 50%). R1, R2, and R3 were the same as in Experiment 2. Thus, R1 had constituent families with as strong a bias as possible towards the linking morpheme *-s-*. R3 showed a bias as strong as possible against *-s-*. The set R2, the neutral set, contained nouns with families without a clear preference for or against *-s-*.

Similar to the previous experiments each of the two sets of left pseudo-constituents (L1, L2) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: Bias in the Left Position (Positive and Negative) and Bias in the Right Position (Positive, Neutral, and Negative). Appendix C lists all experimental items. The 6 x 21 = 126 experimental items were divided over three lists. List 1 contained the compounds of the factorial combinations L1-R1 and L2-R2. List 2 contained the compounds of the combinations L1-R2 and L2-R3, and List 3 contained the compounds of the combinations L1-R3 and L2-R1. In this way, each participant saw each constituent only once. We constructed a separate randomized list of the 2 x 21 = 42 compound constituent pairs for each participant.

Procedure. The procedure was identical to that of Experiments 1 and 2.

Participants. Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch, none had participated in the previous experiments. Each group was asked to complete one of the three experimental lists.

Results and discussion

Occasionally, participants filled in a question mark or a letter sequence other than a linking morpheme. Such responses were counted as errors. The overall error rate was extremely low (0.2%), which allowed us to include all participants and all items

Table 2.3: Percentages of selected linking morphemes when varying Bias for *-s-* in the Left Position (Positive and Negative) and Right Position (Positive, Neutral, and Negative). Standard deviations between parentheses.

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	s	84.0	(14.9)	86.4	(9.0)	79.5	(14.7)
	not s	16.0	(14.9)	13.6	(9.0)	20.5	(14.7)
Negative							
	s	24.8	(17.4)	20.0	(15.4)	16.4	(16.9)
	not s	75.2	(17.4)	80.0	(15.4)	82.6	(16.9)

in the data analysis. Table 2.3 summarizes the percentages of *s* responses versus other responses for the six experimental conditions. Appendix C lists the individual words together with the absolute numbers of *s* and *not s* responses.

A by-item logit analysis of the *s* and *not s* responses revealed a main effect of Bias in the Left Position ($F(1,120) = 276.0$, $p < .0001$), no effect of Bias in the Right Position ($F(2,120) = 2.2$, $p = .1201$), and no interaction of Bias in both positions ($F(2,120) = .6$, $p = .5726$).

Experiment 3 addressed the question whether the family of the preceding suffix affects the choice for or against the linking morpheme *-s-* when building a new nominal compound. We found a strong effect of the suffix family on the choice of the linking morpheme. We were not able to replicate the smaller, but significant effect of the family of the right constituent which we have seen in Experiments 1 and 2. The use of pseudo-words in the Left Position led to compounds which are difficult to interpret. Maybe the lack of a possible interpretation decreased the effect of the bias in the Right Position which was already small in the previous two experiments.

Summary: Experimental results

Experiments 1 and 2 have revealed that linking morphemes in novel compounds can be predicted on the basis of the families of both left and right constituents, and that the effect of the left family is much stronger. We have seen that the difference in strength mirrors a distributional asymmetry in the lexicon, i.e. left constituents tend to have a stronger bias for or against a linking morpheme than right constituents. Experiment 3 has shown that suffixes attached to pseudo-words to form left con-

stituents also affect the choice of linking morphemes.

The experimental results are in line with the descriptions in the literature in so far as the properties of the left constituent are traditionally described as the main factors influencing the choice of linking morphemes. The presence of a weaker, but significant effect of the right constituent is in line with the observation that right constituents may be important because they codetermine the semantic relation between the constituents in a compound. We have also shown that the final suffix in derived left pseudo-words plays a role, which is in line with the observations reported by Van den Toorn (1981a; 1981b) for real words. Most importantly, the results of our experiments have revealed unambiguous evidence for a strong analogical effect of the constituent family, a novel factor that is not discussed in the linguistic literature.

In the next section, we proceed to test whether it is possible to simulate the effect of the constituent families with the help of an explicit computational algorithm for analogy. The aim of this section is to ascertain whether analogy based on constituent families is computationally tractable. In the general discussion, we will outline how the computational technique that we have opted for can be mapped onto a psycholinguistically plausible architecture of the mental lexicon.

Analogical modeling

Several techniques are available for the modeling of data which display statistical tendencies rather than discrete regularities. Connectionist models are widely used to obtain predictions for graded data where standard rule-based methods fail. Although connectionist networks are powerful nonlinear classifiers, they have the disadvantage that additional follow-up analyses of the network are required in order to understand how the network arrives at its classifications. A second disadvantage of connectionist models is that it is at present unclear whether they can accommodate the family size effect reported in Schreuder & Baayen (1997) and De Jong, Schreuder, & Baayen (2000). The family size effect concerns the finding that type counts of morphologically related words for target words correlate with lexical decision times and subjective frequency ratings to these target words, while the corresponding token counts have emerged as irrelevant. Given the sensitivity of connectionist networks to frequencies of occurrence, i.e., token frequencies, it is as yet unclear how this type frequency effect might emerge in combination with the absence of the token frequency effect. As the role of the constituent family that has

emerged from our experiments appears to be a similar type count effect, but now in production rather than in comprehension, we have opted for an exemplar-based approach in which type counts effects are more easily accommodated

Exemplar-based approaches have been developed by, e g , Skousen (1989) and Daelemans, Zavrel, Van der Sloot, & Van den Bosch (1999) Skousen has proposed an analogical model specifically for the domain of language In his model, stored exemplars are compared with a given target word using a similarity metric defined over a series of user-specified features Exemplars that are most similar to the target are most likely to serve as the analogical basis for its classification

Various machine-learning techniques proceed along similar lines We have opted for a program implementing a series of machine-learning techniques, TiMBL, developed by Daelemans et al (1999) [5] This implementation offers powerful heuristics for finding directly the features with a strong analogical weight In what follows, we first describe this machine-learning technique, which we have found very useful from a computational linguistics point of view We then discuss the results that we have obtained with this technique In the general discussion, we outline the way in which the technical computational model can be mapped onto a psycholinguistically more plausible model of analogical processing in the mental lexicon

Exemplar-based learning

Exemplar-based learning techniques implement the idea that the performance of cognitive processes is based on explicit storage of representations of earlier experiences Reasoning is conducted by comparing a new instance with stored instances Crucially, the information carried by earlier experiences is not extracted from these experiences and stored in the form of abstract rules Instead, a general strategy for similarity-based reasoning is combined with the extensive storage of exemplars in an instance database For example, the problem of assigning the position of the main stress to a novel Dutch word is solved by storing large numbers of multi-syllabic words in the instance database, and by using a distance measure defined over the phonological make-up of the final two syllables of these words A search in the instance base leads to the exemplar which is most similar to that of the target noun The stress position stored with this exemplar is suggested to be that of the target noun (see Daelemans, Gillis, & Durieux, 1994, for a detailed study) The main advantage of exemplar-based learning is that no abstract rules

[5] For a detailed comparison between TiMBL and Skousen s AML model see Krott, Schreuder & Baayen (in press, also chapter 3)

need to be formulated. The price to be paid is that computational load increases substantially with the size of the database, because the distance between any new instance and all exemplars in the instance database must be computed. We will return to the issue of this computational load below.

In our experience, the *k*-NN algorithm with the Hamming distance measure known as IB1 in machine learning literature (Aha, Kibler, & Albert, 1991) yields the best results for the modeling of Dutch linking morphemes. Its similarity metric is very simple. Given two patterns *X* and *Y*, each represented by *n* features, the distance between *X* and *Y* is the number of shared features. TiMBL makes three additions to the original *k*-NN algorithm. First, the value of *k* refers to the *k*-nearest distances and not the *k*-nearest cases. In our simulation studies we have set *k* to unity, which means that all instances at Hamming-distance 1 are included in the set of nearest neighbors. Second, if the nearest neighbor set contains more than one instance, the linking morpheme is selected that is most often instantiated in this nearest neighbor set. Third, in case of a tie, the linking morpheme is selected that has the highest frequency in the instance base.

TiMBL has the useful possibility to add to the Hamming-distance measure a relevance weight for every feature (the IB1-IG algorithm). TiMBL accomplishes this by means of the information gain (IG) which looks at a feature and measures how much information it contributes to our knowledge of the correct linking morpheme. The information gain of a feature *i* is obtained by calculating the difference in uncertainty or entropy between the situations without and with knowledge of the value of that feature:

$$\text{IG} = w_i = H(C) - \sum_{v \in V_i} P(v) \cdot H(C|v). \tag{2.1}$$

In (2.1), *C* denotes the set of linking possibilities (*-en-, -s-,* ∅), and V_i the set of values for feature *i* (e.g., 'stressed' and 'unstressed' for the feature Stress). The entropy of the linking possibilities is

$$H(C) = -\sum_{c \in C} P(c) \log_2 P(c), \tag{2.2}$$

with *c* ranging over {*-en-, -s-,* ∅}. Using information gain weights, we get the following distance metric:

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i) \tag{2.3}$$

(Daelemans et al., 1999: 9). By computing the information gain for the many features that one might potentially use in a particular simulation study, it becomes

possible to make an informed preselection of features.

In what follows, we will apply this methodology to the materials of the first two experiments in order to ascertain to what extent machine learning techniques are able to predict the choice of linking morphemes.

Predicting linking morphemes

In order to gauge the predictive power of exemplar-based learning of Dutch linking morphemes, we first studied the preferred choices for existing compounds using 10-fold cross-validation. In 10-fold cross-validation the dataset is divided into 10 'held-out' subsets. For each held-out subset, linking morphemes are predicted on the basis of the remaining 90% of the data, which serve as the training set. The overall performance of the model is evaluated in terms of the average percentage of correctly predicted linking morphemes calculated over the 10 cross-validation runs.

A crucial determinant of the model's performance is the set of features defining its input space. In our simulation studies, we have made use of 9 features. The first and second features code the left and right immediate constituents, which represent the left and right constituent families. The third feature represents the plural suffix selected by the left constituent. This feature can be used to extract the knowledge that the linking morpheme *-en-* is found only after left constituents that select *-en* as their plural suffix. Features 4–7 code the abstractness and animacy of the first and the second constituent. They allow us to trace whether the semantics of the constituents codetermine the choice of linking morphemes (Van den Toorn, 1982a). Feature 8 marks the presence of stress on the final syllable of the first constituent, as it might be possible that the linking morpheme *-en-* is inserted to avoid a stress clash between the two constituents. Finally, feature 9 codes the morphological complexity of the first constituent in terms of its number of morphemes as a greater complexity of the left constituent has been argued to give rise to a preference for *-s-* (see Mattens, 1984). In various simulation runs not reported here, we used the three final phonemes of the first constituent, the three initial phonemes of the second constituent, as well as the last morpheme of the first constituent instead of features 1 and 2. As the results obtained with this alternative feature set invariably turned out to yield inferior results, we do not discuss these alternative features.

We used the 22,994 Dutch nominal compounds in the CELEX lexical database that occur with a frequency of at least 2 per 42 million word forms as our instance

base. Each of these compounds was assigned a vector of values for our 9 features. The second column of Table 2.4 lists the information gain for each individual feature on the basis of the training sets in the cross-validation runs. When we use all features, we predict the correct linking morpheme for 93.2% of the compounds in the held-out datasets. When we use only the first feature, the first constituent, which has the highest information gain, we obtain an accuracy which is only slightly less, 92.5%. The linguistic literature describes the choice of linking morpheme as governed by a conspiracy of tendencies. Our cross-validation results suggest that, indeed, these tendencies allow the linking morpheme to be predicted with a high degree of accuracy. Surprisingly, most of the predictive power resides in a single feature only: the first constituent, i.e., the key for the morphological family of the first constituent.

How well does the model predict the choice of the linking morpheme for the neologisms used in Experiments 1 and 2? First consider Experiment 1 summarized in columns 4–6. The column labelled Fam1 lists information gain and accuracy when the model is trained on pooled constituent families of all experimental words. We trained the model on this subset of the compounds listed in CELEX for the following reason. The semantic specification for a constituent of a given compound, as we have used it for the first study, is not restricted to the meaning of the constituent in this particular compound, but provides the full range of possible meanings the constituent can have when used in isolation. For a specific compound, this range of possible feature values is too broad. For the subset of constituent families it was feasible to manually narrow down the semantics to the correct meaning for each specific compound separately. Consequently, there are two differences between this analysis and the previous analysis based on the CELEX data. First, the semantic features are more precise, second, the number of types on which TiMBL is trained is much smaller (CELEX 22,994 vs. Fam1 1864).

When we train on the pooled families using all features, we obtain an accuracy of 83.6%. As we are dealing with neologisms, accuracy is evaluated in terms of the percentage of experimental words for which TiMBL predicts a linking morpheme that is identical to the majority choice of our participants. Again, we observe that the first constituent has the highest information gain, and that using this feature exclusively already leads to an accuracy of 78.8%. By adding features 5 and 6, we can increase the accuracy to 85.2%. Feature 5 concerns the animacy of the left constituent: Animate left constituents elicit higher numbers of *en* responses. Feature 6 represents the abstractness of the right constituent: Abstract right nouns

Table 2.4: Features used in the simulation studies, their information gain (upper part of the table), and the corresponding prediction accuracy (lower part of the table). Celex: results using 10-fold cross-validation. EN, S: results for Experiments 1 and 2, with accuracy being evaluated against the majority choice of the participants. Predictions are made on the basis of various training sets: Fam1: pooled family members of all experimental items; CELEX: all compounds in CELEX; Fam2: predictions based on left and right constituent families of each individual item. *: features determined as relevant by forward step-wise selection.

No.	Feature	Celex	EN			S		
			Fam1	CELEX	Fam2	Fam1	CELEX	Fam2
1	1st C	1.11*	1.29*	1.11*	*	1.14*	1.11*	*
2	2nd C	0.41	0.96	0.41		0.70	0.41	
3	1st C: plur	0.10	0.12	0.10		0.07	0.10	
4	1st C: abst	0.07	0.13	0.07		0.13	0.07	
5	1st C: anim	0.04	0.13*	0.04	*	0.07	0.04	
6	2nd C: abst	0.02	0.06*	0.02	*	0.06*	0.02*	*
7	2nd C: anim	0.00	0.01	0.00*		0.01	0.00	
8	1st C: stress	0.07	0.13	0.07		0.07	0.07	
9	1st C: compl	0.11	0.05	0.11		0.08	0.11	
	accuracy 1–9	93.2%	83.6%	78.3%	84.7%	91.5%	82.5%	82.5%
	accuracy 1	92.5%	78.8%	75.1%	79.9%	87.8%	82.5%	87.3%
	accuracy *	92.5%	85.2%	82.0%	86.8%	91.5%	83.1%	88.4%

lead to fewer *en* responses. The selection of these features is based on forward step-wise selection. At the first step, the feature with the highest information gain is selected. For each successive step, the feature with the next highest information gain is considered. If addition of this feature improves accuracy, it is added to the list of features. Otherwise, the feature with the next highest information gain is tested. The information gains of the features selected by this algorithm are marked with an asterisk in Table 2.4.

When we compare these results with those obtained with cross-validation for all compounds in CELEX (column 3), we observe a decrease in accuracy of roughly 10%. This loss of accuracy has three possible sources. First, the experiment made use of neologisms, non-existing compounds presented without a natural context, that may have been somewhat more artificial than existing compounds. However, whatever the nature of our materials may be, the performance of the model is similar to that of human subjects. When we calculate the average accuracy of the subjects in the same way as we evaluate the accuracy of the model, i.e., by treating the majority choice as norm, we obtain an average accuracy of 85.1%, which comes close to the maximum of the range of model accuracies (78.8–85.2). Apparently, participants and the model find the task equally difficult.

Second, the set of words with a Neutral Bias in the experiment is atypical for the population as a whole. As we have already seen in Figure 2.1, most of the left constituents in CELEX reveal a strong bias for or against *-en-* (98% of all left constituents appear with the linking morpheme *-en-* either in less than 35% or in more than 65% of all members of the constituent family). The over-representation of left constituents without a strong bias in the experiment (30% versus 2% off all CELEX compounds) renders the experiment more difficult to model than the CELEX population of compounds using cross-validation. In fact, the accuracy scores for the subsets of words with a strong bias for or against *-en-* are substantially higher than those for the words with a Neutral Bias (Left Positive Bias: 92.1%; Left Neutral Bias: 71.4%; Left Negative Bias: 90.5%). Clearly, the atypical Neutral set renders the experiment more difficult.

Third, the reduced size of the training set may have led to reduced accuracy. To investigate this possibility we ran additional simulation experiments. When we train the model on all compounds in CELEX rather than on the subsets of words for which we checked the coding of concreteness and animacy of the constituents by hand, we observe a slight reduction in accuracy of roughly 3%. Possibly, this reduction arises because the semantic coding is less precise for the database as a whole. In-

terestingly, we obtain slightly improved accuracies when we train the model not on a larger but on an even smaller training set. By training on the unique family members of each experimental compound separately, we improve the average accuracy to 86.8% (column 6, Fam2), using the same features that led to the highest accuracy when training on the pooled family members.[6] It is remarkable that training on the basis of small by-item families (with a range of 8-84 family members) results in slightly, although not significantly ($p > .2$, proportions test), improved performance compared to training on the 1864 pooled family members or the 22,994 compounds in CELEX. This suggests that the constituent families provide the analogical basis for selecting the linking morphemes in novel compounds. From a psycholinguistic perspective, this is an important result as it obviates the need to scan the complete lexicon for analogical exemplars. In the general discussion, we shall use this result to formulate a psycholinguistic spreading activation model for the analogical selection of linking morphemes.

The last three columns of Table 2.4 summarize the results obtained using the same procedures for the data of Experiment 2. The best results are obtained when we train TiMBL on the pooled constituent family members of all experimental compounds. On the basis of the first constituent and the abstractness of the second constituent (abstract right constituents lead to more *s* responses), TiMBL achieves an accuracy of 91.5%. When we train the model on the compounds in CELEX, accuracy decreases significantly to 83.1% ($p = .02$, proportions test). Training on the individual families of the experimental compounds leads to a slight reduction in accuracy that, however, does not differ significantly from the accuracy when trained on the pooled constituent family members. Compared to the participants of Experiment 2, who on average opt for the majority choice for 83.5% of the experimental compounds, TiMBL performs surprisingly well.

The results summarized in Table 2.4 are the best results that we have been able to obtain. Replacing the features for the first and second constituents by features for the last three segments of the first constituent and the first three segments of the second constituent invariably leads to decreasing performance. The same holds for training on the last morpheme of the first constituent.

Table 2.5 compares the success rate that can be achieved on the basis of the phonological and morphological rules that have been formulated for Dutch with the

[6]One might expect to achieve the same accuracy for Fam1, Fam2, and CELEX when training only on the first constituent (accuracy 1). However, the different numbers of training items and the resulting different structures of the three TiMBL-internal decision trees as well as the random choice of linking morphemes in the case of ties lead to somewhat different results.

Table 2.5: Comparison of rule-based and analogy-based predictions for experiments 1 and 2. x/y: number of successful prediction/number of applicable cases; phonology: predictions based on the final rime; morphology: predictions based on the final suffix; semantics: predictions based on semantic rules for mass nouns, human agents ending in *-er*, and synthetic compounds in which the left constituent is the non-subject argument of the embedded verb to its right.

	EN (experiment 1)			
	applicable		not applicable	
	rules	TiMBL	rules	TiMBL
phonology	9/15	13/15	-/174	142/174
morphology	15/36	36/36	-/153	119/153
semantics	8/14	10/14	-/175	245/175
	S (experiment 2)			
	applicable		not applicable	
	rules	TiMBL	rules	TiMBL
phonology	12/24	24/24	-/165	133/165
morphology	27/51	41/51	-/138	116/138
semantics	11/34	28/34	-/155	129/155

corresponding success rate as achieved by TiMBL (trained on the constituent families of the the individual items), for experiments 1 and 2. Note that the rules are applicable only to small subsets of the materials. The phonological rules state that no linking morpheme is allowed following a rime ending with a vowel, with a liquid preceding /k/ or /m/, or with a schwa followed by a sonorant. For words with other rime characteristics, the rules provide no predictions at all. Not surprisingly, the morphological rules apply only to the compounds in our materials which have a derived left constituent. Similarly, the semantic rules apply only to words with a mass noun and human agents ending in *-er* as left constituent, as well as to synthetic compounds in which the left constituent is the non-subject argument of the embedded verb to its right. From Table 2.5, it is clear that TiMBL outperforms the rules for all applicable words. In addition, TiMBL provides good predictions where the rules provide none. Interestingly, TiMBL reveals the animacy and abstractness of the left and right constituents to be relevant factors co-determining to some extent the choice of the linking morpheme. Further rigorous quantitative research will have to clarify which semantic factors contribute to the choice of the linking morpheme over and above the constituent families themselves.

Finally, Table 2.6 presents a comparison of the performance of the participants with the performance of TiMBL when trained on the constituent families of the the

Table 2.6: Comparison of the participants and TiMBL across experimental conditions. Number of participants (averaged over items) selecting *-en-* in Experiment 1 and *-s-* in Experiment 2 and the corresponding expectations based on TiMBL (see text).

Left	Right	EN (Experiment 1)		S (Experiment 2)	
		participants	TiMBL	participants	TiMBL
pos	pos	19.0	17.8	19.3	19.7
pos	neutr	19.3	18.3	19.5	19.7
pos	neg	17.5	17.9	18.3	19.7
neutr	pos	15.0	11.3	13.3	10.4
neutr	neutr	16.4	12.4	12.7	10.4
neutr	neg	11.7	11.8	10.5	10.4
neg	pos	3.6	0.0	2.7	0.0
neg	neutr	3.8	0.0	1.0	0.0
neg	neg	1.2	0.0	0.4	0.0

individual items. The first two columns specify the Bias (Positive, Neutral, or Negative) for the left and right constituents. The third and fifth columns list the number of participants (averaged over items) that selected *-en-* (column 3) and *-s-* (column 5) in Experiments 1 and 2 respectively. TiMBL provides for each item the probabilities for the various linking options. Given that there were 20 participants in each of the two experiments, the expected number of participants selecting, e.g., *-en-* in Experiment 1 for a given item equals 20 times the probability of *-en-* for that item. The average number of participants selecting *-en-* for the nine experimental conditions of Experiment 1 and 2 are listed in columns 4 and 6 respectively. Note that the expected values as predicted by TiMBL are similar to the experimental values, and this impression is confirmed by goodness of fit tests.[7] Thus, the predictions of TiMBL as a computational model of analogy remain accurate even when we consider the individual conditions of our experimental design.

Note that this is not a trivial result. The model could have failed in several ways. First, it could have predicted linking morphemes at chance level. This would have indicated that constituent bias would not be the true factor underlying the choice of linking morphemes. In that case, our conclusion would have been that we failed to include the appropriate features in the input data. Second, the model could have

[7]For Experiment 1, $\chi^2_{(8)} = 6.44$, p = .60 and for Experiment 2, $\chi^2_{(8)} = 9.05$, p = .34. In order to avoid technical problems with zero counts for the negative left bias conditions, the chi-squared tests were actually run on the complement counts for all conditions, i.e., the number of participants not selecting *-en-* (Experiment 1) or *-s-* (Experiment 2).

predicted the correct choice for the wrong reasons. Suppose that the model had based its predictions not on the constituent family but on the nature of the third phoneme of the right constituent. Suppose, furthermore, that the left constituent family bias is uncorrelated with the nature of this third phoneme. In these circumstances, the model would be interesting from a technical point of view but seriously flawed from a cognitive point of view, as our experiments show that constituent bias is an important factor if not the most important factor. Third, we ran our simulation studies not only on the bases of the constituent families but on a great many other features as well. The simple fact that the model assigns the greatest information gain to the constituent families is not an artifact of the selection of our experimental materials, as can be seen from the cross-validation data obtained for all noun-noun compounds in the CELEX lexical database.

Summing up, the present simulation studies show that predictions mirroring the actual choices of human participants can be made on the basis of the families of the left constituent in combination with the semantics of both constituents. These results suggest that analogy may well underlie the strong intuitions that language users have concerning the choice of the appropriate linking morpheme.

General discussion

This study has addressed the question of how analogy influences the choice of linking morphemes in Dutch noun-noun compounds. Even though the usage of linking morphemes in noun-noun compounds is not well predictable by rule, it can be quite well predicted analogically on the basis of the constituent families of both the left and the right constituents. It is the family of the left constituent which constitutes the primary domain of analogical prediction for existing words (Experiments 1 and 2). In the case of suffixed pseudo-words as left constituents, the suffix provides the analogical domain for the choice of the linking morpheme (Experiment 3). A series of computational simulation studies using an exemplar-based machine-learning algorithm for the modeling of analogy, TiMBL, revealed that the actual linking morphemes selected by the participants in our experiments can be predicted with a high degree of accuracy on the basis of the morphological family of the first constituent with some additional influence of the semantics of the second constituent. These results lead us to conclude that the left constituent families provide the crucial analogical basis for selecting the most appropriate linking morpheme in Dutch. When comparing the choices made by the participants in our experiments

with those made by the machine-learning algorithm, we found that the selection is equally difficult for human subjects and TiMBL.

Our results show that the choice of the linking morpheme hinges on existing exemplars with the same left constituent. At the same time, our experimental evidence suggests that the right constituent has a minor role to play. We know of three other studies that mention a possible role for the left constituent. For compounds in Afrikaans, Botha (1968) argued that nouns are lexically marked for linking morpheme when they appear as left constituents in compounds. This works fine for those left constituents that consistently occur with only one linking morpheme. However, for the many left constituents with variable realizations, Botha is forced to assume lexical listing of the full compounds. Unfortunately, Botha's theory has no predictive power with respect to neologisms which have a left constituent with variable realizations.

The idea that analogy might be involved has been suggested for German linking morphemes by Becker (1992), who, however, makes use of such a general notion of analogy that it is difficult to see how any falsifiable predictions might be obtained. Dressler, Libben, Stark, Pons, & Jarema (2001) present experimental data that hint at a role for left constituent bias in German, but these authors mention this possibility only in passing for a small subset of their data. Since our present results show that it is possible to explicitly model analogy quantitatively and to predict its influence experimentally, we believe that we now have a realistic methodology for studying the influence of analogy on the realization of linking morphemes across a wider range of languages.

Recall that there is considerable variation in the realization of the linking morphemes. We have seen this variation in the responses of the participants in our experiments, and it is also visible in comprehensive dictionaries, which list variants such as *spelling+wijziging* and *spelling+s+wijziging* ('spelling change') side by side. This variation is captured by our analogical model, which allows for some uncertainty with respect to the appropriate linking morpheme exactly as observed for the responses of our participants. This kind of variation is not restricted to linking morphemes, it is also found in the domain of derivational morphology. For instance, Malicka-Kleparska (1985) discusses the formation of diminutives in Polish and calls attention to the free variation between the rival forms *-ik* and *-ek* that occurs for words with a particular phonological form. Such free variation is at odds with strict rule-based systems, while it may arise in systems based on analogy in the absence of a clear bias for a particular form. We believe that such variational data provide

evidence in favor of the view that morphological rules are grounded in analogy.

Thus far, we have used the machine-learning algorithm implemented in TiMBL to model the analogical selection of linking morphemes in novel compounds. From a computational linguistics point of view, TiMBL captures the analogy underlying the linking morphemes quite satisfactorily. From a psycholinguistics point of view, the question arises whether it is realistic to assume that in general analogy is really based on an exhaustive calculation of a distance metric for all forms in the lexicon. In fact, TiMBL itself does not carry out such an exhaustive calculation for a novel form. While this might be feasible on a massively parallel machine, present-day sequential machines require alternative algorithms. TiMBL solves this algorithmic problem by constructing a decision tree during training (Daelemans, Van den Bosch, & Weijters, 1997). By dropping a novel form through the decision tree, the appropriate linking morpheme is identified.

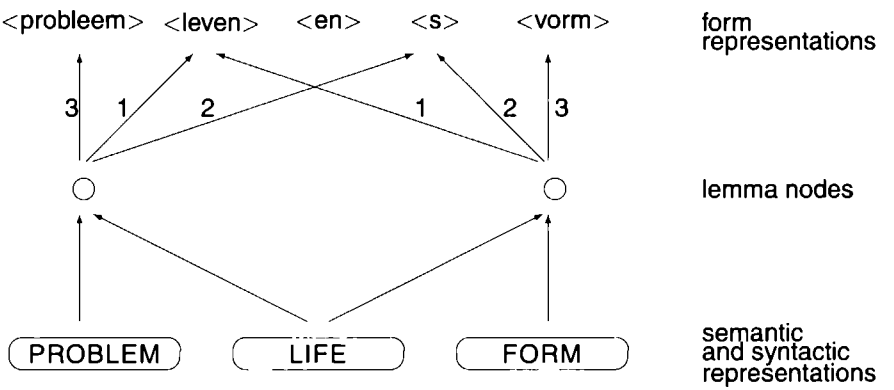


Figure 2.4: Selected semantic representations, lemma nodes, and form representations for the two lexicalized compounds *leven+s+probleem* ('life problem', left lemma node) and *leven+s+vorm* ('life form', right lemma node).

Such a decision tree can in fact be understood as a set of rules. Given that the analogy underlying the choice of linking morphemes is based on constituent families, a separate rule for each constituent is embodied in the decision tree. Those researchers who view morphological processing as fundamentally rule-based therefore have the option of reformulating the decision tree of TiMBL as a set of morphological rules. The cost of this option is a proliferation of rules, one for each possible left constituent. As we find this cost too high, we have explored an alternative ap-

proach based on the idea of parallel co-activation of constituents in a spreading activation framework along the lines of Schreuder & Baayen (1995). Parallel co-activation is a realistic option precisely because our experimental results have revealed that it is only the constituent families that have to be inspected, and not each and every compound in the mental lexicon (or in TiMBL's instance base). Consider Figure 2.4. The units in the bottom layer in Figure 2.4 represent sets of semantic and syntactic features. For instance, the unit labelled PROBLEM is a short-hand representation for a series of syntactic and semantic representations such as NOUN, ABSTRACT, INANIMATE etc. Even though not represented graphically in Figure 2.4, representations such as those for NOUN and ABSTRACT are shared by the units LIFE and FORM. The central layer contains lemma nodes, nodes that link sets of semantic and syntactic representations to form representations. For instance, the left-hand lemma, representing *leven+s+probleem* ('life problem'), is activated during production by the semantic and syntactic representations of PROBLEM and LIFE and in turn activates the form representations <leven>, <probleem>, and <s>. The numbers accompanying the outgoing arrows specify the order in which the form representations have to be linearized for articulation.

In this architecture, the choice for the linking morpheme *-s-* for the novel compound *leven+?+therapie* made by 19 out of 20 participants in Experiment 2 might proceed as follows. Once the syntactic and semantic representations of LIFE and THERAPY have been activated, activation spreads to their lemma nodes. In turn, activation spreads from the lemma nodes to their form representations, activating <leven> and <therapie>. Because *leven+?+therapie* does not have its own lemma representation, and because the linking morphemes are not themselves addressed, the form representations of linking morphemes have not yet been activated.

It has recently been shown that in subjective frequency ratings and in visual lexical decision, morphological families of target words are coactivated (De Jong et al., 2000; Schreuder & Baayen, 1997). Our hypothesis is that in production an analogous coactivation of the constituent families takes place. Thus, we assume that the semantic and syntactic representations for the left constituent LIFE in Figure 2.4 coactivates the lemmas of *leven+s+vorm* ('life form'), *leven+s+probleem* ('life problem') and other such compounds when the target word is *leven+?+ therapie*.[8,9] The

[8]For evidence of storage of regular complex words in Dutch see Baayen, Dijkstra & Schreuder (1997), Bertram, Schreuder & Baayen (2000); for compounds see Van Jaarsveld & Rattink (1988)

[9]It is in principle possible that compounds are activated which contain *leven* as a right constituent as in *student+en+leven* 'student life' However, a post-hoc analysis showed that the family homo-

lemmas of these constituent family members in turn coactivate their form representations, including their linking morphemes.[10]

In addition to the strong influence of the first constituent, we have also seen a somewhat weaker effect of the right constituent in our experiments, both factorially and in the correlation analyses of bias and response. We can model the prominence of the left constituent families by having the semantic and syntactic representations of the left constituent, LIFE in our example, send extra activation to the lemma nodes with which it is connected. Possibly, the special burst of activation flowing from the first constituent to the lemma layer is a consequence of it being the first constituent that has to be articulated (Roelofs, 1996).[11] Recall that the TiMBL results revealed an effect of the semantics of the right constituent. For instance, abstract right constituents show a slight preference for the linking morpheme *-s-*. We assume that right abstract constituents coactivate lemma nodes for abstract nouns, and therefore also abstract noun compounds in the constituent families. The activation of these compound lemma nodes leads to some extra support for the linking morpheme *-s-*.

Finally, the results of Experiment 3, in which the left constituents were suffixed pseudo-words, can be understood along similar lines. Under the assumption that the suffix in the pseudo-word activates its semantics, and that these semantics in turn coactivate the lemmas of the compounds with this suffix, the bias in the suffix family will lead to a preference for a given linking morpheme.

The present results challenge the idea that in order to model non-deterministic linguistic phenomena symbolic representations have to be given up and replaced by subsymbolic representations as argued by, for instance, Rumelhart & McClelland (1986a) and Seidenberg (1987); see also Zhou & Marslen-Wilson (unpublished manuscript). We have shown that it is possible to model analogy without giving up symbolic representations such as lemmas for complex words. At the same

geneity of these compounds in Experiment 2 is not correlated with the response homogeneity. This is true for compounds containing left constituents at the right position (r_s = 18, z = 1 44, p = 15) as well as for compounds containing right constituents at the left position (r_s = 01, z = 04; p = 97) These results suggest that only those family members of the left constituent are activated which share the left constituent with the novel compound, and only those family members of the right constituent which share the right constituent with the novel compound

[10]Figure 2 4 illustrates the composition route of our parallel dual route model We assume that there is also a full-form representation <levens>, the plural of <leven>, for which support can accumulate in the same way as for <s>.

[11]The prominence of the first constituent is in line with the observed greater priming effects of first constituents reported by Kehayia, Jarema, Tsapkini, Perlak, Ralli, & Kadzielawa (1999) In addition, Stark & Stark (1991) report impaired production of second constituents of compounds by a Wernicke's aphasic

time, we do not think it is necessary to be committed to the view that morphological rules are in essence symbolic rewrite-rules. This formal view of word formation rules is challenged by the experimental and simulation results for the compounds with neutral bias that we have studied. Here, both our participants and our model showed great uncertainty with respect to what might be the most appropriate linking morpheme. This uncertainty is difficult to reconcile with formal deterministic rules. For strongly converging, consistent domains, formal analogical models will show behavior similar to that of deterministic rules. For diverging, inconsistent domains, deterministic rules impose regularity that is not present in the data nor, if we may trust our experimental results, in the minds of speakers of Dutch. Formal models of analogy, on the other hand, reflect the inconsistency present in their input domains both in the variation in their output and in the confidence they assign to their output. This shows that formal models of analogy are not unconstrained all-powerful theories that can always predict any outcome and hence have no explanatory value. Instead, the behavior of formal models of analogy is tightly constrained by its input domain. For Dutch compounds, local family-based analogical generalization instead of global lexicon-based rule generalization has allowed us to approximate human behavior with greater precision and insight.

References

Aha, D. W., Kibler, D. and Albert, M.: 1991, Instance-based learning algorithms, *Machine Learning* **6**, 37–66.

Anshen, F. and Aronoff, M.: 1988, Producing morphologically complex words, *Linguistics* **26**, 641–655.

Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **36**, 94–117.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Schreuder, R., De Jong, N. H. and Krott, A.: in press, Dutch inflection: the rules that prove the exception, *in* S. Nooteboom, F. Weerman and F. Wijnen (eds), *Storage and Computation in the Language Faculty*, Kluwer Academic Publishers, Dordrecht.

Becker, T.: 1992, Compounding in German, *Rivista di Linguistica* **4**(1), 5–36.

Bertram, R., Laine, M., Baayen, R. H., Schreuder, R. and Hyönä, J.: 1999, Affixal homonymy triggers full-form storage even with inflected words, even in a morphologically rich language, *Cognition* **74**, B13–B25.

Bertram, R., Schreuder, R. and Baayen, R. H.: 2000, The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Memory, Learning, and Cognition* **26**, 419–511.

Botha, R. P.: 1968, *The Function of the Lexicon in Transformational Grammar*, Mouton, The Hague.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Clahsen, H., Eisenbeiss, S. and Sonnenstuhl-Henning, I.: 1997, Morphological structure and the processing of inflected words, *Theoretical Linguistics* **23**, 201–249.

Cleveland, W. S.: 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

Daelemans, W., Gillis, S. and Durieux, G.: 1994, The acquisition of stress, a data-oriented approach, *Computational Linguistics* **20**(3), 421–451.

Daelemans, W., Van den Bosch, A. and Weijters, A.: 1997, IGTree: Using trees for compression and classification in lazy learning algorithms, *Artificial Intelligence Review* **11**, 407–423.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 1999, TiMBL: Tilburg Memory Based Learner Reference Guide 2.0, *Report 99-01*, Computational Linguistics Tilburg University.

De Haas, W. and Trommelen, M.: 1993, *Morfologisch handboek van het Nederlands* (Morphological handbook of Dutch), SDU, Den Haag.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

De Saussure, F.: 1966, *Course in General Linguistics*, McGraw, New York.

Dressler, W. U., Libben, G., Stark, J., Pons, C. and Jarema, G.: 2001, The processing of interfixed German compounds, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp. 185–220.

Fienberg, S.: 1980, *The Analysis of Cross-Classified Categorical Data*, The MIT Press, Cambridge, Mass.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Halle, M. and Marantz, A.: 1993, Distributed morphology and the pieces of inflection, *in* K. Hale and S. Keyser (eds), *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, Vol. 24 of *Current Studies in Linguistics*, MIT Press, Cambridge, Mass, pp. 111–176.

Kehayia, E., Jarema, G., Tsapkini, K., Perlak, D., Ralli, A. and Kadzielawa, D.: 1999, The role of morphological structure in the processing of compounds: The interface between linguistics and psycholinguistics, *Brain and Language* **68**, 370–377.

Krott, A., Schreuder, R. and Baayen, R. H.: in press, Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds, *in* R. Skousen (ed.), *Analogical Modeling: An Exemplar-Based Approach to Language*.

Kuipers, A. H.: 1960, *Phoneme and Morpheme in Kabardian*, Mouton and Co., The Hague.

Malicka-Kleparska, A.: 1985, Parallel deriviation and lexicalist morphology: the case of Polish diminutivization, *in* E. Gussmann (ed.), *Phono-Morphology. Stud-*

ies in the Interaction of Phonology and Morphology, Catholic University of Lublin, Lublin, pp. 95–112.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R. and Pinker, S.: 1995, German inflection: The exception that proves the rule, *Cognitive Psychology* **29**, 189–256.

Marle, J. v.: 1985, *On the Paradigmatic Dimension of Morphological Creativity*, Foris, Dordrecht.

Mattens, W. H. M.: 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343.

Neijt, A., Baayen, R. H. and Schreuder, R.: in preparation, Reading relicts of the past: The semantics of linking elements in present-day Dutch orthography.

Pinker, S.: 1991, Rules of language, *Science* **153**, 530–535.

Pinker, S.: 1997, Words and rules in the human brain, *Nature* **387**, 547–548.

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Rietveld, T. and Van Hout, R.: 1993, *Statistical Techniques for the Study of Language and Language Behaviour*, Mouton de Gruyter, Berlin.

Roelofs, A.: 1996, Serial order in planning the production of successive morphemes of a word, *Journal of Memory and Language* **35**, 854–876.

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S. and Mars, F.: 1997, Morphological priming, fragment completion, and connectionist networks, *Journal of Memory and Language* **36**(3), 382–405.

Rumelhart, D. E. and McClelland, J. L. (eds): 1986, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, MIT Press, Cambridge, Mass.

Sandra, D., Frisson, S. and Daems, F.: 1999, Why simple verb forms can be so difficult to spell: the influence of homophone frequency and distance in Dutch, *Brain and Language* **68**(1/2), 277–283.

Schreuder, R. and Baayen, R. H.: 1995, Modeling morphological processing, *in* L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum, Hillsdale, New Jersey, pp. 131–154.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Schreuder, R , Neijt, A , Van der Weide, F and Baayen, R H 1998, Regular plurals in Dutch compounds linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573

Seidenberg, M 1987, Sublexical structures in visual word recognition Access units or orthographic redundancy, *in* M Coltheart (ed), *Attention and Performance XII*, Lawrence Erlbaum Associates, Hove, pp 245–263

Seidenberg, M and Hoeffner, J 1998, Evaluating behavioral and neuroimaging data on past tense processing, *Language* **74**, 104–122

Sereno, J and Jongman, A 1997, Processing of English inflectional morphology, *Memory and Cognition* **25**, 425–437

Skousen, R 1989, *Analogical Modeling of Language*, Kluwer, Dordrecht

Spencer, A and Zwicky, A (eds) 1998, *The Handbook of Morphology*, Blackwell, Oxford

Stark, J and Stark, H -K 1991, On the processing of compound nouns by a Wernicke's aphasic, *in* J Tesak (ed), *Neuro- und Patholinguistik*, Vol 35 of *Grazer Linguistische Studien*, Universitat Graz, Graz, pp 95–112

Taft, M 1979, Recognition of affixed words and the word frequency effect, *Memory and Cognition* **7**, 263–272

Van den Toorn, M C 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is (Linking phonemes in compounds with derived forms as first constituents), *De Nieuwe Taalgids* **74**, 197–205

Van den Toorn, M C 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is (Linking phonemes in compounds with loanwords as first constituents), *De Nieuwe Taalgids* **74**, 547–552

Van den Toorn, M C 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I)), *De Nieuwe Taalgids* **75**(1), 24–33

Van den Toorn, M C 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for the regulation of linking phonemes in nominal compounds II), *De Nieuwe Taalgids* **75**(2), 153–160

Van Jaarsveld, H and Rattink, G 1988, Frequency effects in the processing of lexicalized and novel nominal compounds, *Journal of Psycholinguistic Research* **17**, 447–473

Zhou, X and Marslen-Wilson, W 1999, Lexical representation of compound

words: cross-linguistic evidence, *Unpublished manuscript*.

Appendices

Appendix A

Materials for Experiment 1: left constituent and right constituent (number of *en* responses, number of other responses).

L1-R1: Left Position: Positive *-en-* Bias; Right Position: Positive *-en-* Bias:
student kolder (20, 0); pen prik (20, 0); advocaat geslacht (18, 2); soldaat deken (19, 1); vreemdeling buurt (20, 0); kleur tegenstelling (10, 10); sigaret knipsel (18, 2); sigaar kiosk (17, 3); pan rook (19, 1); toerist klooster (20, 0); roos gaas (20, 0); beer lever (20, 0); noot laan (18, 2); aap klauw (20, 0); tomaat moes (20, 0); kat haat (19, 1); reus hol (20, 0); gans lijf (20, 0); stier beet (20, 0); vrucht massa (20, 0); wesp ras (20, 0)

L1-R2: Left Position: Positive *-en-* Bias; Right Position: Neutral *-en-* Bias:
noot dief (19, 1); sigaret bundel (20, 0); sigaar republiek (17, 3); stier kooi (20, 0); kat paar (20, 0); wesp jacht (20, 0); aap vel (19, 1); vrucht rek (20, 0); tomaat stam (20, 0); roos zee (19, 1); soldaat bond (20, 0); pen hout (20, 0); gans boter (20, 0); kleur rad (19, 1); student kas (20, 0); reus rijk (20, 0); beer galerij (17, 3); pan kaas (15, 5); vreemdeling steun (20, 0); toerist kuil (20, 0); advocaat corps (20, 0)

L1-R3: Left Position: Positive *-en-* Bias; right constituent: Negative *-en-* Bias:
sigaar juffrouw (20, 0); sigaret tarief (20, 0); tomaat project (18, 2); pan lengte (11, 9); toerist gedeelte (20, 0); soldaat bevoegdheid (17, 3); beer maaltijd (19, 1); aap terrein (20, 0); vreemdeling crisis (20, 0); student voorschrift (20, 0); gans schade (18, 2); advocaat weg (17, 3); kleur techniek (13, 7); noot gewas (11, 9); pen patroon (12, 8); vrucht kanaal (18, 2); roos kunst (20, 0); kat therapie (17, 3); wesp deskundige (19, 1); reus vrijheid (19, 1); stier psycholoog (18, 2)

L2-R1: Left Position: Neutral *-en-* Bias; Right Position: Positive *-en-* Bias:
begrip tegenstelling (7, 13); bloem laan (20, 0); bom massa (14, 6); bron gaas (11, 9); buur geslacht (15, 5); god hol (13, 7); heer buurt (20, 0); kaart kiosk (20, 0); koe ras (18, 2); klas kolder (19, 1); kool moes (8, 12); leerling klauw (13, 7); lid lijf (10, 10); persoon beet (7, 13); pijp rook (11, 9); plaat knipsel (19, 1); pop klooster (19, 1); prul deken (19, 1); wolf lever (12, 8); woord haat (20, 0); ziel prik (20, 0)

L2-R2: Left Position: Neutral *-en-* Bias; Right Position: Neutral *-en-* Bias:
begrip stam (11, 9); bloem boter (14, 6); bom kuil (16, 4); bron rijk (15, 5); buur steun (19, 1); god vel (11, 9); heer kaas (18, 2); kaart bundel (20, 0); klas republiek (20, 0); koe kooi (20, 0); kool rek (16, 4); leerling corps (19, 1); lid kas (15, 5); persoon bond (13, 7); pijp galerij (18, 2); plaat hout (11, 9); pop rad (19, 1); prul zee (19, 1); wolf paar (14, 6); woord jacht (19, 1); ziel dief (17, 3)

L2-R3: Left Position: Neutral *-en-* Bias; Right Position: Negative-*en-* Bias:
begrip patroon (10, 10); bloem weg (20, 0); bom lengte (11, 8); bron terrein (13, 7); buur project (12, 7); god maaltijd (16, 4); heer tarief (18, 2); kaart juffrouw (18, 2); koe psycholoog (17, 3); kool gewas (3, 17); leerling bevoegdheid (12, 8); lid voorschrift (11, 9); persoon therapie (3, 17); pijp schade (4, 16); plaat techniek (11, 9); pop kunst (14, 6); prul kanaal (12, 8); ziel vrijheid (6, 14). woord gedeelte (3, 17); klas crisis (19, 1); wolf deskundige (12, 8)

L3-R1: Left Position: Negative *-en-* Bias; Right Position: Positive *-en-* Bias:
stad haat (2, 18); gevangenis deken (0, 20); neus knipsel (6, 14); angst prik (2, 18); industrie rook (4, 16); wijn kiosk (4, 16); kalf beet (2, 18); bevolking ras (0, 20); bier lever (8, 12); overheid geslacht (0, 20); christen klooster (6, 14); dokter klauw (0, 20); fabriek buurt (4, 16); dak gaas (5, 15); aardappel moes (3, 17); rivier massa (15, 5); citroen laan (10, 10); groep hol (0, 20); wetenschap kolder (1, 19); kwaliteit tegenstelling (3, 17); koning lijf (1, 19)

L3-R2: Left Position: *-en-* bias; Right Position: Neutral *-en-* Bias:
stad republiek (0, 20); industrie corps (7, 13); bevolking stam (0, 20); dokter bond (2, 18); rivier hout (8, 12); dak kuil (6, 14); groep jacht (3, 17); kwaliteit kaas (0, 20); angst steun (6, 14); aardappel bundel (5, 15); wijn dief (2, 18); kalf kooi (4, 16); koning vel (1, 19); bier zee (5, 15); neus paar (17, 3); wetenschap rijk (0, 20); overheid kas (0, 20); gevangenis rek (3, 17); citroen boter (3, 17); christen galerij (6, 14); fabriek rad (1, 19)

L3-R3: Left Position: Negative *-en-* Bias; Right Position: Negative *-en-* Bias:
aardappel juffrouw (2, 18); angst crisis (1, 19); bevolking gedeelte (0, 20); bier deskundige (1, 19); christen vrijheid (2, 18); citroen gewas (1, 19); dak lengte (4, 16); fabriek psycholoog (0, 20); gevangenis terrein (0, 20); groep bevoegdheid (0, 20); industrie weg (1, 19); kalf maaltijd (4, 16); koning therapie (2, 18); kwaliteit

kunst (0, 20); neus kanaal (2, 18); overheid project (0, 20); rivier techniek (5, 15); stad patroon (0, 20); wetenschap voorschrift (0, 20); wijn schade (0, 20)

Appendix B

Materials for Experiment 2: Left constituent and right constituent (number of *s* responses, number of other responses).

L1-R1: Left Position: Positive *-s-* Bias; Right Position: Positive *-s-* Bias:
arbeider standpunt (20, 0); bedrijf bevoegdheid (19, 1); beslissing angst (19, 1); bestuur aangelegenheid (20, 0); fabriek norm (20, 0); gezicht dimensie (16, 4); groep afstand (19, 1); handel fractie (20, 0); investering orientatie (20, 0); leven tactiek (19, 1); macht woede (18, 2); onderzoek reden (20, 0); ontwikkeling duur (20, 0); persoonlijkheid bevordering (20, 0); regering verhouding (20, 0); staat besluit (19, 1); training toename (19, 1); veiligheid drang (20, 0); verkeer delegatie (20, 0); verzorging bijdrage (20, 0); vrede uitoefening (18, 2)

L1-R2: Left Position: Positive *-s-* Bias; Right Position: Neutral *-s-* Bias:
arbeider functie (20, 0); bedrijf organisatie (20, 0); beslissing conflict (18, 2); bestuur regel (20, 0); fabriek geschiedenis (19, 1); gezicht verandering (19, 1); groep plicht (18, 2); handel project (20, 0); investering kunst (20, 0); leven therapie (19, 1); macht dienaar (20, 0); onderzoek niveau (20, 0); ontwikkeling patroon (20, 0); persoonlijkheid controle (20, 0); regering kwaliteit (20, 0); staat conferentie (16, 4); training probleem (20, 0); veiligheid mechanisme (20, 0); verkeer rust (20, 0); verzorging commissie (20, 0); vrede karakter (20, 0)

L1-R3: Left Position: Positive *-s-* Bias; Right Position: Negative *-s-* Bias:
arbeider tent (20, 0); bedrijf bos (15, 5); beslissing schrift (13, 7); bestuur club (19, 1); fabriek kaas (20, 0); gezicht tekening (17, 3); groep kast (15, 5); handel voorraad (19, 1); investering meester (20, 0); leven bel (20, 0); macht laag (19, 1); onderzoek schaal (19, 1); ontwikkeling sprong (19, 1); persoonlijkheid spiegel (19, 1); regering les (20, 0); staat eiland (13, 7); training olie (20, 0); veiligheid venster (20, 0); verkeer soort (19, 1); verzorging transport (20, 0); vrede stok (19, 1)

L2-R1: Left Position: Neutral *-s-* Bias; Right Position: Positive *-s-* Bias:
begrip dimensie (14, 6); bisschop fractie (17, 3); directeur besluit (19, 1); dood re-

den (18, 2); generaal delegatie (13, 7); geschut afstand (19, 1); geweld bijdrage (20, 0); god woede (10, 10); heil bevordering (15, 5); klimaat verhouding (11, 9); lucifer norm (9, 11); minister bevoegdheid (12, 8); monnik aangelegenheid (0, 20); persoon angst (14, 6); plicht uitoefening (17, 3); president standpunt (12, 8); temperatuur toename (14, 6); tijd orientatie (16, 4); voordracht duur (18, 2); voorkeur drang (20, 0); wolf tactiek (8, 12)

L2-R2: Left Position: Neutral *-s-* Bias; Right Position: Neutral *-s-* Bias:
begrip probleem (12, 8); bisschop karakter (18, 2); directeur commissie (12, 8); dood rust (16, 4); generaal functie (19, 1); geschut mechanisme (15, 5); geweld organisatie (14, 6); god dienaar (11, 9); heil therapie (11, 9); klimaat geschiedenis (10, 10); lucifer kwaliteit (6, 14); minister plicht (11, 9); monnik regel (6, 14); persoon kunst (17, 3); plicht verandering (18, 2); president conferentie (12, 8); temperatuur controle (15, 5); tijd conflict (19, 1); voordracht niveau (17, 3); voorkeur patroon (17, 3); wolf project (8, 12)

L2-R3: Left Position: Neutral *-s-* Bias; Right Position: Negative *-s-* Bias:
begrip laag (10, 10); bisschop spiegel (18, 2); directeur stok (14, 6); dood eiland (9, 11); generaal kast (18, 2); geschut tent (12, 8); geweld soort (10, 10); god bos (5, 15); heil olie (9, 11); klimaat schaal (7, 13); lucifer voorraad (3, 17); minister club (14, 6); monnik kaas (6, 14); persoon transport (6, 14); plicht schrift (4, 16); president bel (9, 11); temperatuur venster (14, 6); tijd sprong (14, 6); voordracht les (13, 7); voorkeur tekening (18, 2); wolf meester (12, 8)

L3-R1: Left Position: Negative *-s-* Bias; Right Position: Positive *-s-* Bias:
avond duur (5, 15); boek bijdrage (0, 20); christen aangelegenheid (0, 20); dak afstand (5, 15); dwang reden (0, 20); kleur verhouding (5, 15); licht dimensie (5, 15); morgen delegatie (3, 17); nacht tactiek (2, 18); natuur bevordering (6, 14); nood besluit (3, 17); slag uitoefening (1, 19); soldaat woede (3, 17); straat orientatie (1, 19); student standpunt (0, 20); vuur angst (2, 18); wapen bevoegdheid (3, 17); wijn norm (3, 17); woning fractie (4, 16); zand toename (2, 18); zang drang (4, 16)

L3-R1: Left Position: Negative *-s-* Bias; Right Position: Neutral *-s-* Bias:
avond functie (1, 19); boek organisatie (0, 20); christen commissie (0, 20); dak controle (1, 19); dwang regel (1, 19); kleur kwaliteit (0, 20); licht kunst (1, 19); morgen rust (1, 19); nacht project (0, 20); natuur therapie (0, 20); nood mechanisme (1, 19);

slag niveau (0, 20); soldaat dienaar (3, 17); straat karakter (0, 20); student conflict (0, 20); vuur patroon (0, 20); wapen geschiedenis (3, 17); wijn conferentie (0, 20); woning verandering (9, 11); zand probleem (1, 19); zang plicht (0, 20)

L3-R1: Left Position: Negative *-s-* Bias; Right Position: Negative *-s-* Bias:
avond sprong (0, 20); boek transport (0, 20); christen schrift (1, 19); dak kast (0, 20); dwang soort (0, 20); kleur schaal (0, 20); licht spiegel (0, 20); morgen bos (2, 18); nacht tent (0, 20); natuur eiland (0, 20); nood olie (0, 20); slag les (0, 20); soldaat stok (2, 18); straat bel (1, 19); student kaas (0, 20); vuur venster (0, 20); wapen club (0, 20); wijn laag (0, 20); woning tekening (2, 18); zand voorraad (0, 20); zang meester (0, 20)

Appendix C

Materials for Experiment 3: Left constituent and right constituent (number of *s* responses, number of other responses).

L1-R1: Left Position: Positive *-s-* Bias; Right Position: Positive *-s-* Bias:
ontbolfing aangelegenheid (18, 2); verbrimming afstand (18, 2); bebuiping angst (18, 2); wouking besluit (12, 8); hernabbeling bevoegdheid (18, 2); struffing bevordering (18, 2); snoking bijdrage (15, 5); bronkheid delegatie (20, 0); golheid dimensie (19, 1); pritsheid drang (20, 0); dulligheid duur (20, 0); sloefheid fractie (19, 1); spreunheid norm (19, 1); vlitheid orientatie (18, 2); conviriteit reden (15, 5); descaliteit standpunt (10, 10); dipromeniteit tactiek (14, 6); illuniteit toename (15, 5); recarveniteit uitoefening (18, 2); solutaniteit verhouding (18, 2); virubaniteit woede (11, 9)

L1-R2: Left Position: Positive *-s-* Bias; Right Position: Neutral *-s-* Bias:
ontbolfing commissie (18, 2); verbrimming conferentie (18, 2); bebuiping conflict (18, 2); wouking controle (15, 5); hernabbeling dienaar (18, 2); struffing functie (16, 4); snoking geschiedenis (12, 8); bronkheid karakter (19, 1); golheid kunst (18, 2); pritsheid kwaliteit (16, 4); dulligheid mechanisme (19, 1); sloefheid niveau (20, 0); spreunheid organisatie (19, 1); vlitheid patroon (18, 2); conviriteit plicht (16, 4); descaliteit probleem (16, 4); dipromeniteit project (19, 1); illuniteit regel (17, 3); recarveniteit rust (17, 3); solutaniteit therapie (18, 2); virubaniteit verandering (16, 4)

L1-R3 Left Position Positive *-s-* Bias, Right Position Negative *-s-* Bias
ontbolfing bel (20, 0), verbrimming bos (18, 2), bebuiping club (13, 7), wouking eiland (10, 10), hernabbeling kaas (12, 8), struffing kast (11, 9), dipromeniteit laag (17, 3), vlitheid les (19, 1), golheid meester (19, 1), pritsheid olie (15, 5), dulligheid schaal (19, 1), sloefheid schrift (19, 1), spreunheid soort (17, 3), bronkheid spiegel (18, 2), conviriteit sprong (16, 4), descaliteit stok (14, 6), snoking tekening (13, 7), illuniteit tent (15, 5), recarveniteit transport (14, 6), solutaniteit venster (17, 3), virubaniteit voorraad (18, 2)

L2-R1 Left Position Negative *-s-* Bias, Right Position Positive *-s-* Bias
moepsel aangelegenheid (4, 16), lirksel afstand (3, 17), steukster angst (9, 11), raalster besluit (7, 13), vilkster bevoegdheid (7, 13), girdin bevordering (4, 16), kloerdin bijdrage (3, 17), dreekster delegatie (14, 6), preuksel dimensie (3, 17), pleefster drang (11, 9), veepsel duur (3, 17), taapster fractie (8, 12), brumsel norm (4, 16), zwaagster orientatie (7, 13), borberin reden (1, 19), doerin standpunt (0, 20), darsin tactiek (5, 15), stimsel toename (2, 18), vlatsel uitoefening (5, 15), ploebin verhouding (2, 18), zwaperin woede (2, 18)

L2-R2 Left Position Negative *-s-* Bias, Right Position Neutral *-s-* Bias
taapster commissie (6, 14), girdin conferentie (1, 19), raalster conflict (8, 12), preuksel controle (0, 20), ploebin dienaar (3, 17), steukster functie (10, 10), stimsel geschiedenis (5, 15), dreekster karakter (5, 15), pleefster kunst (7, 13), veepsel kwaliteit (2, 18), vlatsel mechanisme (2, 18), moepsel niveau (2, 18), vilkster organisatie (8, 12), lirksel patroon (1, 19), borberin plicht (3, 17), doerin probleem (0, 20), darsin project (6, 14), zwaagster regel (9, 11), kloerdin rust (3, 17), zwaperin therapie (2, 18), brumsel verandering (1, 19)

L2-R3 Left Position Negative *-s-* Bias, Right Position Negative *-s-* Bias
steukster bel (11, 9), lirksel bos (2, 18), pleefster club (12, 8), zwaperin eiland (1, 19), kloerdin kaas (1, 18), zwaagster kast (6, 14), vlatsel laag (4, 16), dreekster les (5, 15), veepsel meester (3, 17), raalster olie (4, 15), moepsel schaal (1, 18), taapster schrift (7, 13), vilkster soort (2, 18), brumsel spiegel (3, 17), borberin sprong (0, 20), doerin stok (0, 19), darsin tekening (3, 17), girdin tent (0, 20), stimsel transport (3, 17), ploebin venster (1, 19), preuksel voorraad (0, 20)

Analogical hierarchy

CHAPTER 3

This chapter will be published as Andrea Krott, Robert Schreuder, and R. Harald Baayen: Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds. In: Royal Skousen (ed.): Analogical Modeling: An Exemplar-Based Approach to Language.

Abstract

This study compares two exemplar-based models, AML and TiMBL, with respect to their performance in simulating a partly non-deterministic morphological phenomenon, the choice of the linkers *-en-*, *-s-*, and *-∅-* in Dutch noun-noun compounds. We present experimental evidence that the feature selection for the analogical process underlying this choice adapts to the information which is available in the target compound. The three main relevant features, the first constituent of the compound, the suffix, and the rime of the first constituent, are selected on the basis of a fall-back strategy. We also present experimental results which suggest that these three features are hierarchically ordered. The feature Constituent provides the strongest predictor. Its influence overrules the influence of the Suffix and Rime. The feature Suffix in its turn overrules the influence of the Rime. Independent evidence for the hierarchy is provided by the increase of participants' uncertainty when the choice is based on a lower-ranked feature. Simulation studies of the participants' responses in all experiments with AML and TiMBL resulted in excellent fits to the experimental data. AML and TiMBL almost always reach the same high degree of prediction accuracy. Comparing the uncertainty in the models' predictions reveals that these models do not differ in their prediction uncertainty.

Introduction

Traditionally, formal rewrite rules are understood as the normal way to create novel words, while analogy is taken as an unformalizable and exceptional way to create a new word on the basis of an existing word (see e.g., Anshen & Aronoff, 1988). The rule-based approach appears to be adequate for phenomena with strong systematicities which can be easily captured by deterministic rules. However, the very same phenomena can often be described equally well by means of formal and computational models of analogy. In the analogical approach, all novel words are modeled on one or more similar existing forms which serve as the analogical set. Especially in the case of gradual phenomena, where rules often capture only the more or less deterministic sub-patterns in the data, the rule-based approach becomes unsatisfactory. It is these phenomena above all which form a testing ground for the two kinds of approaches.

One of these gradual phenomena is the use of linkers in Dutch noun-noun compounds. There are two main linkers, *-en-*[1] (e.g., *boek+en+kast*, book+LINK+shelf, 'book shelf') and *-s-* (*dame+s+fiets*, woman+LINK+bike, 'woman's bike'), which are historically case endings. Synchronically, they are still homographic with the two nominal plural suffixes. Nevertheless, there are two reasons why it is inaccurate to describe them as plural markers. First, the linking *-s-* occurs in compounds in which it does not form a plural with the first constituent (e.g., *schaap+s+kooi* sheep+LINK+stable 'sheepfold'; the plural of *schaap* is *schaap+en*). Second, the linking *-en-*, though being always the appropriate plural suffix of the first constituent, does not always contribute plural meaning (e.g., *pan+en+koek* pan+LINK+cake 'pancake').

The majority of noun-noun compounds in Dutch do not contain any linker (e.g., *tand+arts* tooth+doctor 'dentist'). Such compounds resemble English compounds. Nevertheless, linkers appear in 35% of all Dutch compounds in the CELEX lexical database (Baayen, Piepenbrock, & Gullikers, 1995) and their distribution is difficult to predict. On the one hand, there are some deterministic patterns. For instance, *bevolking*, when it is used as a first constituent in a compound, always occurs with the linking *-s-*. CELEX lists 30 compounds with *bevolking* as left constituent, all of which are followed by the linker *-s-* (e.g., *bevolking+s+aantal* population+LINK+number 'number of population'). On the other hand, there is rampant

[1]The *-en-* has an orthographic variant *-e-* which, in standard Dutch, does not differ in pronunciation.

unpredictable variation. The left constituent *getal* 'number' occurs in CELEX equally often with *-s-* (3 times), *-en-* (4 times), and *-∅-* (3 times). An examination of CELEX shows that 8.6% of all first constituents are variable in their combination with linkers. These variable first constituents account for 25% of all CELEX compounds.

Rule-based approaches to the description of the distribution of Dutch linkers (see, e.g., Van den Toorn, 1981a; 1981b; 1982a; 1982b; Mattens, 1984; Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997) list phonological, morphological, and semantic factors. An example of a phonological rule is the claim that first constituents ending in a full vowel are never followed by any linker (e.g., Van den Toorn, 1982a; 1982b; Haeseryn et al., 1997). This rule is not without exceptions, as the example *pygmee+en+volk* pygmy+LINK+people 'pygmy people' shows. Morphologically constraints on linkers are based on preferences of suffixes that appear at the end of first constituents. For instance, the diminutive suffix *-tje* always appears with the linking *-s-* (e.g., *kapper+tje+s+saus*, caper+diminutive suffix+LINK+sauce, 'caper sauce'). In contrast, the suffix *-heid* (similar to English '-ness') usually occurs with *-s-*, but also with *-∅-* and *-en-*. One of the semantic rules claims that linkers never follow mass nouns (e.g., *papier+handel* paper+trade 'paper trade'). This is not true for *tabak* 'tabacco' which always appears with *-s-* (e.g., *tabak+s+rook*, tabacco+LINK+smoke, 'tabacco smoke'). There are also attempts to explain linkers by the syntactic relation between the two constituents. If the first constituent is the logical object of the second constituent, a linking element seems to be absent (counterexample: *weer+s+verwachting*, weather+LINK+forecast, 'weather forecast'). Given the large number of exceptions, Van den Toorn prefers the use of the term 'tendencies' rather than 'rules'. Combining all phonological and morphological rules described in the literature[2], and applying them to the compounds in the CELEX database, we find that they only apply to 51% of all the noun-noun compounds. Moreover, they correctly predict only 63% of the linkers in these compounds, which amounts to only 32% of all CELEX compounds. For a list of all applicable rules see Appendix D. Thus, these rules do not sufficiently describe the distribution of Dutch linkers.

In an earlier study, we show that linkers can be predicted with a high degree of accuracy on the basis of analogy (Krott, Baayen, & Schreuder, 2001, also chapter 2). This study revealed strong evidence that the choice of linkers in novel compounds is determined by the distribution of linkers in the set of stored compounds sharing

[2]We did not test any semantic rules because CELEX does not provide the required semantic information.

the first or second constituent with the novel compound. We will refer to this set as the constituent family. We also demonstrated that in the case of compounds with suffixed pseudo-words as first constituents, the analogical set contains all compounds which share the same final suffix of the first constituent. We will refer to this set as the suffix family. In addition to this experimental evidence, the study also showed that the exemplar-based model TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2000) can predict the choices of the participants in off-line production experiments with a high degree of accuracy (ca. 80%). Rules, however, were available just for a small subset of the produced compounds and they were clearly outperformed by TiMBL.

The first goal of the present study is to come to grips with the problem of feature selection. The experiments reported by Krott et al. suggest that different analogical sets are used depending on the input. In the case of novel compounds with existing first constituents, the selection is based on the constituent family. In the case of novel compounds with suffixed pseudo-words as first constituents, the suffix family is relevant. What happens if the first constituent is a pseudo-word which does not contain a suffix? Possibly, the analogical set for monomorphemic pseudo-words is based on the rime of the pseudo-word. We will refer to this analogical set as the rime family and we will test its influence in Experiment 1.

If constituents, suffixes and rimes of first constituents individually influence the choice for linkers, the question arises whether these three factors are equally important. TiMBL provides for each feature that is used for the analogical prediction an information gain measure (IG) which quantifies how much information the feature contributes to the knowledge of the correct linker. When taking all compounds with derived nouns as first constituents and comparing the features Constituent and Suffix in terms of their information gain, it turns out that the feature Constituent has the highest IG value (1.1), while the feature Suffix has a value of 0.8. The feature with the next highest information gain (0.75) is the Rime of the first constituent. The order of IG values suggests a hierarchy in which the Constituent is a stronger factor than the Suffix, while the Suffix is a stronger factor than the Rime.

The second goal of this study is to empirically verify this Constituent-Suffix-Rime hierarchy. This hierarchy implies that lower-ranked features are effective only when higher-ranked features are absent. We present results of experiments which test the precedence of the constituent over the suffix (Experiment 2) and the rime (Experiment 3), as well as the precedence of the suffix over the rime (Experiment 4).

The third goal of this study is to compare the two state-of-the-art exemplar-based

analogical models, AML (Skousen, 1989) and TiMBL Daelemans, Zavrel, Van der Sloot, and Van den Bosch, 2000) with respect to classification accuracy and prediction uncertainty. We will do this by testing how well these models predict the Dutch compounds in the CELEX lexicon as well as the responses of the participants to Dutch novel compounds in our experiments. We will also compare the uncertainty of participants with the uncertainty of the models.

In what follows, we first describe simulation studies which model the linkers of existing Dutch compounds using AML and TiMBL. These simulation studies show that the feature 'constituent' is the best predictor of linkers, although the features 'suffix' and 'rime' are both strong predictors as well.

In the subsequent section, we present results of simulation studies in which the prediction accuracies of both models are tested for novel compounds. We refer to results of previous experiments which test the influence of the first constituent and the suffix of the first constituent on the choice of the linker. We continue with presenting Experiments 1–4 and the corresponding simulation studies with AML and TiMBL.

Predicting existing compounds

In this section, we test how well AML and TiMBL predict the linkers in existing Dutch noun-noun compounds attested in the CELEX lexical database. For these studies, CELEX compounds with a token frequency of zero in a corpus of 42 million words are not included. Ten-fold cross-validation simulation runs over the remaining 22,966 compounds using different analogical sets led to the results summarized in Table 3.1. The column Feature lists the different sets of features determining the analogical sets. The columns TiMBL and AML list the prediction accuracies for these sets. The rows Constituent, Suffix, and Rime list the percentage of correctly classified CELEX compounds if the model's training and classification is based on the analogical set of the first constituent, the suffix and the rime of the first constituent respectively. The constituent family provides the strongest analogical set which correctly classifies about 92% of the compounds in CELEX.[3] This is an extremely high percentage compared to the 32% that are correctly classified by

[3]All results of TiMBL (version 3.0) in this paper are obtained by using the standard IB1 algorithm, the overlap similarity metric with information gain weighting, and one nearest neighbor for extrapolation. In our simulation studies, this set of parameters has been proven to lead to the best results. For AML we excluded '=' as a variable, set the option 'given' to 'exclude', the option 'probability' to unity, and used the option 'squared' without specifying any frequency range.

Table 3.1: Classification accuracies when training is based on the features Constituent, Suffix, and Rime for both TiMBL and AML. '*' marks the classification accuracy when the training is based only on the 3836 first constituents actually ending in a suffix. '+' marks a significant difference in classification accuracy between TiMBL and AML, evaluated by means of a χ^2 test.

	Accuracy %		
Feature	TiMBL	AML	
Constituent	92.6	92.2	
Suffix	74.6 (92.1)*	74.6 (91.3)*	
Rime	78.2	75.6	+
Rime + Suffix	79.5	76.7	+
Rime + Suffix + Constituent	93.4	92.8	+

the phonological and morphological rules reported in the linguistic literature. Apparently, the rule-based approach lacks an extremely important factor. However, when AML and TiMBL have to classify the compounds on the basis of the suffix or on the basis of the rime of the first constituent, they already reach an accuracy of 74.6-78.2%, which suggests that phonological and morphological factors are strong predictors as well. If the simulation is restricted to compounds that indeed contain a final suffix, then a classification on the feature Suffix leads to an accuracy as high as 92.3%. Clearly, among the compounds ending in suffixes, the suffix family is an extremely strong predictor. Combining features for the analogical basis generally leads to better results than a classification which is based on only one feature. The row labeled Rime+Suffix lists the results if the models are trained on the rime and the suffix of the first constituent simultaneously. In this case, AML and TiMBL correctly classifies up to 79.5% of all CELEX compounds. The row labeled Rime+Suffix+Constituent shows the results if all three features are combined. This combination leads to the highest classification accuracies of 93.4% (TiMBL) and 92.8% (AML), which are significantly higher than the accuracies reached by training on only the constituent (TiMBL: $\chi^2_{(1)} = 11.08$, $p < .001$; AML: $\chi^2_{(1)} = 5.80$, $p = .016$).

Comparing the classification accuracies of TiMBL and AML, we find that the models perform equally well as long as the classification is based on the first constituent or the suffix of the first constituent (Constituent: $\chi^2_{(1)} = 2.57$, $p = .11$; Suffix, trained on first constituents ending in a suffix: $\chi^2_{(1)} = 9.58$, $p = .21$). Training on the rime family, however, leads to a significant higher accuracy for TiMBL than for AML ($\chi^2_{(1)} = 43.53$, $p < .001$). This is also true for simulations in which the feature Rime is combined with other features (Rime + Suffix: $\chi^2_{(1)} = 52.46$, $p < .001$; Rime + Suffix

+ Constituent: $\chi^2_{(1)} = 6.36$, $p = .01$).

Summing up, classifying existing Dutch compounds on the basis of the analogical sets of the first constituent, the suffix or the rime of the first constituent, leads to surprisingly high percentages of correct classifications. However, the features are quite different in strength. The first constituent seems to be the strongest predictor, followed by the rime and the suffix. The best result has been obtained with the combination of all three features. A comparison of AML and TiMBL revealed that the models perform equally well as long as the classification is not based on the rime of the first constituent.

Predicting novel compounds

In this section, we test how well AML and TiMBL can predict linking elements that were chosen by participants for novel compounds. We summarize two previous studies in which we observed the influence of the constituent family and the suffix family (Krott et al., 2001, also chapter 2). We also present a new experiment which provides evidence for the influence of the rime family. Simulation studies with AML and TiMBL reveal that these analogical models accurately predict the choices of the linkers made by the participants. Both models reach about the same level of prediction accuracy.

Constituent and Suffix influence

Krott et al. (2001, also chapter 2) tested the influence of the distribution of linkers in the constituent family in two experiments in which participants had to form novel compounds from two visually presented nouns. The first experiment focused on the use of the linking *-en-* (EN-experiment), the second on the use of the linking *-s-* (S-experiment). Both experiments tested the influence of the left and right constituent family. The left constituent family was defined as the set of compounds which share the left constituent with the novel target compound, and the right constituent family was defined as the set of compounds which share the right constituent with the target compound. Constituents for the target compounds were chosen such that the distribution of linkers in the left as well as in the right constituent families varied in their bias for the linker *-en-* (EN-experiment) and *-s-* (S-experiment). The bias was defined as the percentage of compounds in the constituent family which contain *-en-* (or *-s-*). The responses of the participants in both

experiments showed a strong effect of the bias of the left constituent family and a weaker, but still reliable effect of the bias of the right constituent family. The strength of the bias for a linker was positively correlated with the number of responses with this linker.

Krott et al. also present simulation studies in which the responses of the participants were modeled with using TiMBL as analogical model. Because of the variation of the responses for each experimental compound, the prediction of TiMBL was compared with the majority choice of the participants for each compound. Using the constituent family of the first constituent, TiMBL correctly predicted 75.1% of all compounds of the EN-experiment and 82.4% of all compounds of the S-experiment. Modeling the responses with AML leads to results which do not differ significantly from the results obtained with TiMBL (EN-experiment: 82.5%, $\chi^2_{(1)}$ = 2.68, p = .10; S-experiment: 82.0%, $\chi^2_{(1)} < 1$). The results of both models do not change if the analogical set is based on the Constituent, the Suffix and the Rime. Thus, the constituent family seems to provide the main analogical basis.

Krott et al. also investigated whether the suffix of the first constituent influences the choice of the linker, in an experiment in which all first constituents were pseudo-words ending in suffixes. The families of these suffixes differed in their bias for the linking *-s-*. Participants appear to be sensitive to this bias and used the linking *-s-* significantly more often in the case of a strong bias for *-s-* than in the case of a strong bias against *-s-*.

The choices of linkers for the experimental compounds can again be simulated by AML and TiMBL. If we base the classification on the suffix family, the models correctly predict 70.6% of the majority choices of all compounds of the experiment. This does not change if the rime is included into the feature set.

We have seen that the first constituent and the suffix of the first constituent are both affecting the choice of linkers in novel compounds. AML and TiMBL support these results in predicting the choices of the participants with a high degree of accuracy, using the analogical sets of the constituent family and the suffix family. The prediction accuracies of both models do not differ significantly.

Experiment 1: Rime influence

In this section, we focus on the question whether the choices for linkers in novel Dutch compounds also depend on another feature with a high information gain, the rime of the first constituent. If the first constituent is a pseudo-word and does not contain any suffix, we assume that participants use the rime family to choose the

linker In addition to the experiment, we will test whether AML and TiMBL are again capable of simulating the experimental results

Method

Materials We constructed three sets of 24 phonotactically legal Dutch pseudo-words (L1, L2, L3) to be used as left constituents L1 consisted of pseudo-words with rimes which occur in CELEX most often with a linker Of these pseudo-words, 12 ended in *-an* (there are 117 compounds in CELEX ending in *-an*, 65 0% of which have a linker) and 12 ended in *-eid* (254 compounds, 99 6% with linker) Conversely, L3 consisted of pseudo-words ending in rimes which show a bias against being combined with a linker Of these pseudo-words, 6 ended in *-el* (553 compounds, 86 3% without linker), 6 in *-em* (36 compounds, 97 2% without linker), 6 in *-ij* (158 compounds, 89 9% without linker), and 6 in *-a* (237 compounds, 100% without linker) The neutral set L2 consisted of pseudo-words with rimes showing neither a bias for or against a combination with a linker Of these pseudo-words, 8 ended in *-en* (613 compounds, 52 0% with, 48 0% without linker), 8 in *-oe* (25 compounds, 44 0% with, 56 0% without linker), and 8 in *-ap* (28 compounds, 25 0% with, 75 0% without linker) Each pseudo-word was bi-syllabic Word stress was indicated on the first syllable by using capital letters To exclude a possible influence of an existing word, we made sure that none of the pseudo-words ended in an existing Dutch word

We combined each pseudo-word with an existing right constituent which can appear with all three linking possibilities (*-s-*, *-en-*, and *-∅-*) This resulted in a factorial design with one factor with three levels Rime Bias of the first constituent (Positive, Neutral, and Negative) Appendix A lists all $3 * 24 = 72$ experimental compounds We constructed a separate randomized list for each participant

Procedure The participants performed a cloze-task The experimental list of items was presented to the participants in written form Each line presented a pair of compound constituents separated by two underscores We asked the participants to combine these constituents into new compounds and to specify the most appropriate linker, if any, at the position of the underscores, using their first intuitions We told the participants that they were free to use *-en-* or *-e-* as spelling variants of the linker *-en-* The experiment lasted approximately 10 minutes

Participants Twenty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment All were native speakers of Dutch

Results and discussion

For one compound, one participant filled in a question mark. This response was counted as an error. Figure 3.1 displays the number of responses of linkers (+LINKER) and of no linkers (-LINKER) for the three experimental conditions: Positive (POS), Neutral (NEU), and Negative (NEG) Rime Bias. The number of responses are also listed in Appendix A. As can be seen from this figure, a Positive Rime Bias for using a linker leads to more responses with a linker than a Neutral or Negative Bias. A by-item logit analysis (see, e.g., Rietveld & Van Hout, 1993; Fienberg, 1980) of the responses with a linker versus responses without a linker revealed a main effect of the Rime Bias of the first constituent ($F(2,69) = 22.2$, $p < .0001$). We can therefore conclude that the rime of the first constituent affects the choice of the linker. Participants responded to a Negative Bias surprisingly often with a linker. The Negative Rime Bias seems to be less effective. This is remarkable, since the rimes in this condition have been reported as imposing strong restrictions against the usage of linkers in Dutch in the linguistic literature (see, e.g., Van den Toorn, 1982a; 1982b; Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997). As we will see later, a bias against using a linker seems to be easy to violate in general.

In contrast to the experiments which tested the effect of the constituent and suffix family, participants found this experiment extremely difficult to perform. This suggests that the phonological rules listed in the literature are not as strong as assumed and may in fact have no reality for at least some of our participants.[4] The difficulties with this experiment cannot be due to a weaker strength of the bias because in all experiments the bias in the positive and negative condition was equally strong (EN-experiment: Mean Positive Bias: 91%, Mean Negative Bias: 100%; S-experiment: Mean Positive Bias: 98.7%, Mean Negative Bias: 100%; Suffix Experiment: Mean Positive Bias: 91.9%, Mean Negative Bias: 83.3%; Rime Experiment: Mean Positive Bias: 82.3%, Mean Negative Bias: 93.3%).

Given the difficulties experienced by the participants to complete the task, the uncertainty in their choices (with marginally higher majority choices) does not come as a surprise. Interestingly, AML's and TiMBL's performance with respect to the effect of the Rime Bias reveals a high degree of uncertainty as well. Both models correctly predict about half of the majority choices if they are trained on the rime of the first constituents of the 22,966 CELEX compounds (TiMBL: 47.9%; AML:

[4]Vance (1980) reports similar findings in his study of Lyman's law which predicts the occurrence of rendaku in Japanese compounds. He concludes that rendaku is psychologically real only for a rather small minority of speakers.

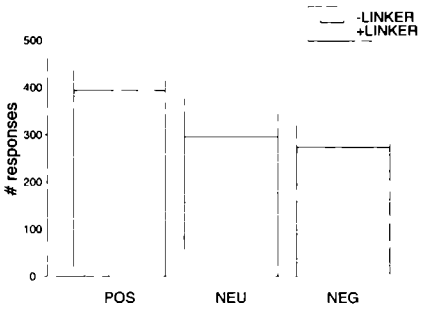


Figure 3.1: Number of responses with linkers (+LINKER) and without linkers (-LINKER) out of a total of 480 responses for the Positive, Neutral, and Negative Rime Bias (POS, NEU, NEG).

47.2%); $\chi^2_{(1)} = 0$, p = 1). However, prediction accuracies increase (TiMBL: 64.8%; AML: 65.3%; $\chi^2_{(1)} = 0$, p = 1) if the training is based not only on the rime but also on the stress of the last syllable of the first constituent.

If the feature set contains Rime, Stress, and Suffix of the first constituent, TiMBL's accuracy drops to 53.4%, while AML's accuracy stays the same with 65.3% ($\chi^2_{(1)} = 1.82$, p = .18). The lower accuracy of TiMBL is due to its analogical mechanism which can lead to level interference of factors. When training is conducted on Rime and Suffix simultaneously, derived and monomorphemic words build separate analogical sub-bases. Consequently, generalizations based on rimes can no longer take priority for the whole dataset.[5]

The uncertainty of choosing linkers

In all AML and TiMBL simulation studies presented in this paper, we investigate how well these models predict the linkers in novel compounds, comparing the linker to which the models assign the highest probability value with the linker which has been chosen most often by the participants. That means that both the less probable linkers for the models and the linkers which are chosen less often by the participants are not taken into account when evaluating the models' performance. In this

[5]Using different parameter settings does not enhance performance. Training on constituent, suffix, and rime while using the IGTREE algorithm leads to 30.1% correctly predicted compounds. With TRIBL we reach 37.0%. If we enhance the number of nearest neighbors for extrapolation to three, both IG and TRIBL reach a prediction accuracy of 30.1%.

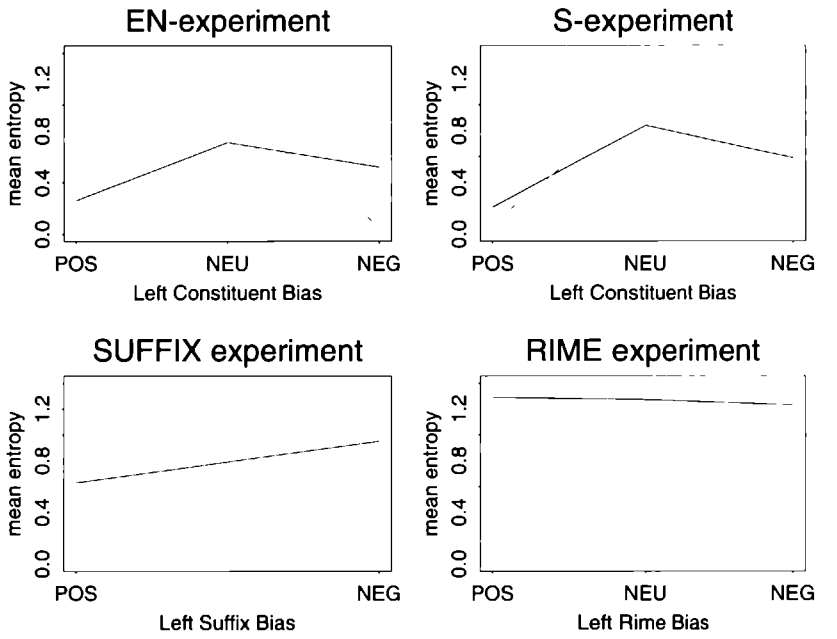


Figure 3.2: Mean entropy for the distributions of choices for linkers for both the models (superimposed dashed lines) and the participants (solid lines) for the experiments testing the influence of the Left Constituent Bias (EN-experiment and S-experiment), the Suffix Bias (SUFFIX experiment), and the Rime Bias (RIME experiment).

section, we will focus on the uncertainty in choosing a linker both on the side of the models and on the side of the participants, addressing the question whether the participants and the models are unsure or sure about the linkers for the same kinds of compounds.

We measured for each compound a model's uncertainty in terms of the entropy of the distribution of the probabilities the model assigns to the linkers *-en-*, *-s-*, and *-∅-* for this compound. The entropy value is the higher the more equally distributed the linkers are. Similarly, we measured the uncertainty of the participants in terms of the entropy of the distribution of the probability values of their choices for a given compound.

Figure 3.2 shows the entropy for different Left Biases in the experiments testing the influence of the Constituent Bias, the Suffix Bias, and the Rime Bias. The upper left panel shows the mean entropy for the three Left Bias conditions in the EN-experiment (Positive, Neutral, and Negative Constituent Bias). The solid line

represents the mean entropy of the distribution of the participants' responses over all experimental items in the three conditions of Left Bias As can be seen from the slope of the line, the entropy, and therefore the uncertainty, is highest in the case of a Neutral Left Constituent Bias This is also true for the entropy of the distributions of the predictions given by the models A Spearman correlation test revealed a significant correlation between the entropy of the participants' responses and the entropy of the models' predictions (r_s = 30, z = 4 14, p < 0001) Interestingly, for this and the following experiments, AML and TiMBL reveal exactly the same average entropy per bias condition

The upper right panel of Figure 3 2 shows the mean entropy for the three Left Bias conditions in the S-experiment Here again, both the models and the participants are most uncertain in the condition of a Neutral Constituent Bias, and the entropy values of the models' predictions and the participants' responses are significantly correlated (r_s = 48, z = 6 79, p < 0001)

Surprisingly, in both the EN-experiment and S-experiment, the models are much more certain in their predictions than the participants for the condition in which the constituent family of the left constituent has a Negative Bias (EN-experiment t(124) = 8 68, p < 0001, S-experiment t(124) = 7 19, p < 0001) There are two explanations for this result First, in the EN-experiment, participants responded most often with *-en-* (2254 out of 3778, $\chi^2_{(1)}$ = 281 33, p < 0001) and in the S-experiment, they responded most often with *-s-* (2092 out of 3780, $\chi^2_{(1)}$ = 85 93, p < 0001) Thus, there might be an overall bias for using *-en-* or *-s-* Second, in the condition of a Left Negative Bias, either 50% (EN-experiment) or 90% (S-experiment) of the left constituents have a bias for *-∅-* Post-hoc analyses revealed that a bias against using a linker can be violated more easily than a bias for *-en-* or *-s-* In the EN-experiment, 75% of the responses followed the bias if it was for *-∅-*, while 93 2% followed the bias if it was for *-en-* or *-s-* ($\chi^2_{(1)}$ = 11 06, p < 001) In the S-experiment, 82 4% of the responses followed the bias if it was for *-∅-*, while 93 5% followed the bias if it was for *-en-* or *-s-* ($\chi^2_{(1)}$ = 4 78, p = 003) These results suggest that the *-∅-* linker might not have the status of a morpheme A bias for *-∅-* would then not be positive evidence for a zero-morpheme, but rather negative evidence against using a linker Such negative evidence might be weaker as an analogical factor than positive evidence for *-en-* or *-s-* Participants would then follow the negative bias less often, leading to greater uncertainty about the choice of the appropriate linker

The lower left panel of Figure 3 2 shows the mean entropy for the two Suffix Biases (Positive and Negative) in the experiment testing the influence of the Suffix

Bias The models are in general less uncertain about the choices than the participants (t(124) = -5 29, p < 0001) Possibly, using the analogical set of the suffix family is already more difficult than using the constituent family There is again a significant correlation between the entropy values of the participants' choices and the models' predictions (r_s = 30, z = 3 37, p < 001)

As mentioned above, participants found the experiment in which we tested the influence of the Rime Bias extremely difficult to perform Not surprisingly, the entropy values of the participants' responses, shown in the lower right panel of Figure 3 2, are very high Interestingly, the entropy does not differ across the three different conditions (Positive versus Neutral Bias t(46) = 66, p = 52, Positive versus Negative Bias t(46) = 95, p = 35) There is also no correlation between the entropy values of the participants' responses and the models' predictions (r_s = - 10, z = - 80, p = 42) Interestingly, the models are as uncertain in the condition of a Positive Bias as in the condition of a Neutral Bias (t(46) = - 12, p = 90) This uncertainty is probably due to the quite low bias (65%) for half of the compounds in this condition However, most of the responses do follow the bias (82%) The high degree of uncertainty of participants in the condition of a Negative Bias can be again explained by the general observation that a bias for -Ø- can be easily violated

In sum, we have seen that participants and models tend to be uncertain especially in the condition of a neutral bias In all experiments, a negative bias reveals higher uncertainty on the side of the participants than on the side of the models We explained this result by the observation that a bias against using a linker seems to be more easily violated This finding suggests that an analogical model for predicting human performance needs to weight zero-realizations differently than other realizations

The feature hierarchy

The experiments testing the influence of the first constituent, of the suffix, and of the rime have revealed that all three features are effective factors This does not mean, however, that these factors are equally effective under the same conditions Participants may activate the constituent family when it is available If the first constituent does not have a constituent family, participants base their choice on either the suffix family or on the rime family of the first constituent In the case of a derived first constituent, the suffix is involved, while in the case of a pseudo-word without any suffix, the rime is crucial We may be dealing with a fall-back strategy In the

absence of a higher-order unit, the next lower unit determines the analogical set However, what happens if the information given in the input allows the selection of more than one feature? Are all such features activated simultaneously and do they equally affect the choice of the linker? The different information gains which are provided by TiMBL suggest the hypothesis that the features are ordered in strength The influence of the constituent might be stronger than that of the suffix, while in turn the influence of the suffix might be stronger than that of the rime We will test these hypotheses in the following three experiments (Experiment 2–4), and we will use AML and TiMBL to investigate the possible role of different analogical sets

Experiments 2 and 3: Constituent Preference

Experiments 2 and 3 test whether the first constituent has a stronger influence on the choice of linkers than the suffix (Experiment 2) or the rime (Experiment 3) of the first constituent

Experiment 2: Constituent Preference or Suffix Preference

Method

Materials For this experiment, we selected a set of 14 derived nouns whose suffixes are mostly combined with the linking *-s-* (mean 86 8%, *-ing* 91 4%, *-ling* 80 9%, *-eling* 86 7%, *-er* 84 1%) At the same time, these derived nouns, when used as left constituents in compounds, tend to occur without the linker *-s-* (mean 91 7%, range 75 0% – 100%, 10 had a bias for *-∅-* and 4 had a bias for *-en-*) To make sure that the bias for *-en-*, *-s-*, and *-∅-* was equal over the list of experimental items, we added 10 monomorphemic nouns with a bias for *-s-* (mean 98 1%, range 83 3% – 100%) and 6 monomorphemic nouns with a bias for *-en-* (mean 91 1%, range 66 7% – 100%), resulting in 30 left constituents The 10 monomorphemic nouns with a bias for *-s-* served as experimental items for Experiment 3

In order to avoid an influence of the right constituent, we combined these 30 left constituents with right constituents which appear with all three linking possibilities (*-s-*, *-en-*, and *-∅-*) Appendix B lists the 16 experimental items We constructed a separate randomized list for each participant

Procedure The procedure was identical to the one used in Experiment 1

Participants Twenty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment All were native speakers of Dutch

Figure 3.3: Number of responses (total: 280) predicted by the Constituent Bias and Suffix Bias.

Results and discussion

None of the participants' responses had to be counted as an error. The left bar of Figure 3.3 shows the number of responses that follow the bias of the constituent, the right bar shows the number of responses that follow the bias of the suffix. The number of responses for the individual compounds are listed in Appendix B. Participants responded most often with the linker that one would expect if they follow the bias of the constituent. Only in 28.6% of all responses, the linker was in line with the bias of the suffix. A paired t-test revealed that Constituent Bias reliably overrides Suffix Bias ($t(13) = 3.04$; $p < .01$). This is especially remarkable considering the fact that a third of the constituents had a bias for -ϕ- that, as we have seen, can be easily overruled. We conclude that the first constituent has indeed a stronger effect on the choice of the linker than the suffix of the constituent.

Simulation studies with TiMBL and AML confirm this result. When we train TiMBL and AML on the first constituents of the 22,966 CELEX compounds, they both correctly predict 64.3% of the majority choices for each experimental compound. If the training is based on the suffix, they correctly predict only 21.4%. Training on the rime, the suffix and the first constituent simultaneously leads to the same results as training on only the first constituent. Therefore, it seems to be mainly the first constituent and its constituent family which is actively used by the participants.

In the next section we address the question whether the bias of the constituent family also overrules the bias of the rime family.

Experiment 3: Constituent Preference or Rime Preference

Method

Materials. We selected from CELEX a set of 10 monomorphemic nouns which tend to occur with a linker (mean: 84.4%; range: 66.7% – 100%). At the same time, the rimes of these nouns tend to occur without a linker (mean: 90.6%; *-ee*: 97.1%; schwa+*l*: 87.9%; *-ij*: 90.6%). Six of these nouns had a bias for a combination with the linker *-en-* and four had a bias for *-s-*. To make sure that the bias for *-en-*, *-s-*, and *-∅-* was equal over the list of experimental items, we added ten derived nouns with a bias against using a linker (mean: 93.9%; range: 63.6% – 100%), four derived nouns with a 100% bias for *-en-*, and six monomorphemic nouns with a 100% bias for *-s-*, resulting in 30 left constituents.

In order to avoid an influence of the right constituent, we combined these 30 left constituents with right constituents which appear with all three linkers (*-s-*, *-en-*, and *-∅-*). Appendix B lists the 10 experimental compounds. We constructed a separate randomized list for each participant.

Procedure. The procedure was identical to the one that was used in Experiments 1 and 2.

Participants. Twenty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch.

Results and discussion

None of the participants' responses had to be counted as an error. The left bar of Figure 3.4 shows the number of responses that follow the bias of the constituent, the right bar shows the number of responses following the bias of the rime. The number of responses of the individual compounds are listed in Appendix B. Figure 3.4 shows that participants responded mostly with the linker following the bias of the constituent. Only in 11.5% of all responses, the linker was in line with the prediction based on the Rime Bias. A paired t-test by items on the number of participants following the bias of the constituent and the number of participants following the bias of the rime confirms that the observed pattern is reliable ($t(9) = 8.6$, $p < .001$). One might put forward that the rime bias is weaker because it is a bias for *-∅-*. However, the difference between the number of responses following the constituent bias (177) and the number of responses following the rime bias (23) is remarkably big. Even if there is a tendency that a bias for *-∅-* can be easier overruled than a bias for *-s-* or *-en-*, this tendency is not strong enough to fully explain the observed difference in responses. Recall that a negative bias in both the EN- and

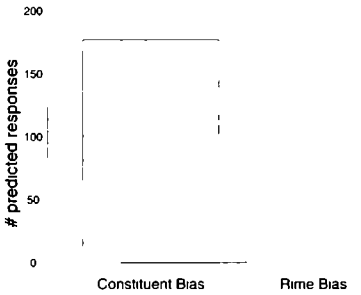


Figure 3.4: Number of responses (total: 200) predicted by the Constituent Bias and Rime Bias.

S-experiments, i.e. a positive bias for *-∅-*, led to a significant decrease of responses with a linker. In addition, a recent study testing the effect of the constituent bias on the choice of linkers in German noun-noun compounds showed that the linker *-∅-* is similarly affected by the constituent bias than the linkers *-(e)n-* and *-s-* (Krott, Schreuder, Baayen, & Dressler, submitted, also chapter 6). We therefore conclude that the influence of the first constituent has a stronger effect on the choice of the linker than the rime of the constituent.

When we train TiMBL and AML on the first constituents of the 22,966 CELEX compounds, both correctly predict 10 out of 10 of the majority choices for each experimental compound. However, when we train on the rime, they correctly predict 0 out of 10. Training TiMBL on the rime, the suffix and the first constituent simultaneously leads to the same results as training on only the first constituent, namely 100% correct predictions. AML's prediction accuracy in this case drops to 90%, which is not significantly lower ($\chi^2_{(1)}$ = .002, p = .96). Clearly, participants base their choices on the constituent family and not on the rime family. In the next section, we will test whether the Suffix Bias is stronger than the Rime Bias.

Experiment 4: Suffix Preference

Method

Materials. We constructed a list of $4 * 3 = 12$ phonotactically legal Dutch pseudowords which ended in 4 different Dutch suffixes which also appear as word-final letter combinations in monomorphemic nouns (*-er*, *-aar*, *-ing*, and *-ist*). When these

letter combinations appear in monomorphemic nouns, they are usually not combined with a linker (mean 72 9%, range 59 4% – 93 5%) In contrast, when they appear as suffixes, they tend to be combined with a linker (mean 84 0%, *-er* 84 1% with *-s-*, *-aar* 66 7% with *-s-*, *-ing* 91 4% with *-s-*, *-ist* 93 8% with *-en-*)

In order to balance the bias for linkers in the experiment, we also constructed 24 filler constituents Half of these were phonotactically legal Dutch derived pseudo-words ending in suffixes that appear always without any linker (*-sel*, *-te*, *-atie*, and *-nis*, 3 pseudo-words for each suffix) The other half of the filler items were phonotactically legal Dutch monomorphemic pseudo-words ending in letter combinations that usually appear with a linker (mean 63 7%, *-eid* 86 0%, *-ap* 37 1%, and *-an* 67 9%, 4 pseudo-words for each combination) For both the 12 experimental items and the 24 fillers, stressed syllables were marked by capital letters

We constructed two lists of experimental items (List A, List B) Both lists contained the 12 experimental pseudo-words To List A we added the 12 filler words which usually appear with a linker To List B we added the 12 fillers which usually appear without a linker Each pseudo-word was embedded in a sentence constructed to influence the interpretation of the pseudo-word For the words of List A, the sentences promoted a monomorphemic interpretation of the pseudo-word For the words of List B, the sentences promoted an affixal interpretation The following two examples show one of the experimental pseudo-compounds preceded by the two contexts

A monomorphemic interpretation
Een 'PLOEver' is een boomsoort *PLOEver_gried*
"A 'PLOEver' is a kind of tree PLOEver_gried"

B derived interpretation
Iemand die graag 'ploeft' is een 'PLOEver' *PLOEver gried*
"Somebody who likes to 'ploef' is a 'PLOEver' PLOEver gried"

In addition, we constructed 12 + 24 = 36 compounds each using a pseudo-word of Lists A and B as a left constituent and combining it with a right phonotactically legal pseudo-word that does not appear in Lists A and B By using right pseudo-constituents, we avoided any additional effect on the selection The compounds with the 12 experimental left constituents were identical in Lists A and B Appendix A lists all sentences and compounds of List A and List B We constructed a separate randomized list for each participant

Procedure The participants performed a cloze-task The sentences defining the pseudo-words and the compounds were presented to the participants in written form Each line presented a sentence and the pair of compound constituents in which the first constituent was identical to the defined pseudo-word The constituents were separated by two underscores The participants were instructed to first read the sentence twice in order to understand the meaning of the pseudo-word Then they had to combine the two constituents into a new compound and to specify the most appropriate linker, if any, at the position of the underscores, using their first intuitions We told the participants that they were free to use *-en-* or *-e-* as spelling variants of the linker *-en-* The experiment lasted approximately 5 minutes

Participants Forty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment All were native speakers of Dutch List A was presented to half of participants, List B to the other half

Results and discussion

All participants provided a linking choice for all items The left bar of Figure 3 5 (derivational interpretation) shows the number of responses when the sentence favors a derivational interpretation As can be seen from the figure, this condition mainly led to responses as predicted by the bias of the suffix (mean 71 5%) The right bar of Figure 3 5 (monomorphemic interpretation) shows the number of responses when the sentence favors a monomorphemic interpretation The number of responses for the individual compounds are listed in Appendix C Paired t-tests of the number of responses for the two contexts show that participants responded more often with the predicted linker for a derived first constituent for a sentence favoring a derivational interpretation than for a sentence favoring a monomorphemic interpretation ($t(11) = 4\ 5$, $p < 001$) They also responded more often with the predicted linker for a monomorphemic first constituent for a sentence favoring a monomorphemic interpretation than for a sentence favoring a derived interpretation ($t2(11) = 2\ 9$, $p = 01$) However, even in the case of a sentence favoring a monomorphemic interpretation, more responses are predicted by the bias of the suffix than by the bias of the rime ($t(11) = 3\ 5$, $p = 004$)

These results lead to two conclusions First, rimes and suffixes of first compound constituents independently influence the choice of linkers Second, the influence of the suffix is much stronger It is the prominent factor even when the pseudo-word is introduced contextually as a monomorphemic word

When we train TiMBL and AML on the suffix of the first constituents of the 22,966

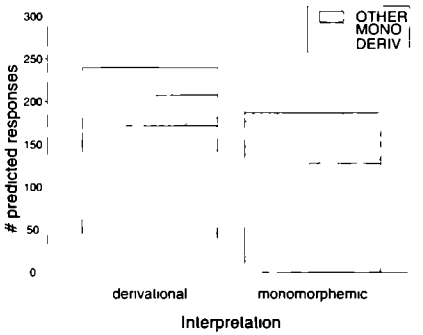


Figure 3.5: Number of responses predicted by the Suffix Bias (DERIV) or the Monomorphemic Bias (MONO), and number of other responses (OTHER) in the two experimental conditions of a derivational and monomorphemic interpretation.

CELEX compounds, they correctly predict 100% of the majority choices for each experimental compound in the case of a preceding sentence favoring a derived interpretation. Their prediction accuracy drops to 83.3% in the case of a preceding sentence favoring a monomorphemic interpretation. When we train the models on the rime instead, they predict only 50% in the case of a sentence favoring a derived interpretation. Their prediction rises to 58.3% in the case of a sentence favoring a monomorphemic interpretation.

These results support the experimental finding that the behavior of the participants is influenced by the context. Participants base their choices more often on the analogical set of the rime instead of the suffix if the preceding sentence favors a monomorphemic interpretation. The results also mirror the stronger influence of the suffix which seems to easily activate the corresponding suffix family when it is present in the input, even when the monomorphemic interpretation of the pseudo-word should inhibit this activation.

General discussion

This study aimed for three goals. First, we tried to come to grips with the problem of feature selection in the task of choosing the appropriate linkers in Dutch noun-noun compounds. Second, we tested whether the three main relevant features for this task, Constituent, Suffix, and Rime, are hierarchically ordered. Third, we simulated the choices of participants with AML and TiMBL and compared these models with

respect to their classification accuracies and their prediction uncertainty

The first goal, solving the problem of feature selection, has been addressed by simulation studies focusing on existing compounds in CELEX and experiments with novel compounds Both kinds of studies have shown that the constituent family, the suffix family, and the rime family all affect the choice of linkers in compounds However, the three factors are not effective to the same extent and under the same conditions The simulation studies with existing compounds revealed that the constituent family seems to provide the strongest analogical set The suffix family is as strong as the constituent family, but only for compounds with first constituents which indeed end in a suffix Otherwise, it is the least effective of the three factors The experiments with novel compounds suggest that the features Suffix and Rime only affect the selection when the next higher-ranked feature (Constituent or Suffix) is absent in the input

At the bottom of the feature hierarchy, the rime family emerges as a rather problematic analogical set Participants reported extreme difficulties in the experiment testing the influence of the rime These difficulties were confirmed by the analyses of the uncertainty in the responses of the participants, which revealed a high entropy across all conditions of this experiment Due to this uncertainty, AML and TiMBL reach a rather low prediction accuracy of maximal 65 3% (AML) and 64 8% (TiMBL), which is considerably less than the accuracies for the experiments testing the influence of the suffix (TiMBL 92 1%, AML 75 4%) and constituent (EN-experiment TiMBL 75 1%, AML 82 5%, S-experiment TiMBL 82 4%, AML 82 0%) Apparently, choosing linkers on the basis of the rime of the first constituent is an unusual task This is not so surprising, considering the fact that for normal compounds there is usually a constituent family or at least a suffix family available which can serve as the analogical set

The second main question of this study was whether the features Constituent, Suffix, and Rime are hierarchically ordered We found experimental evidence suggesting that the Constituent Bias indeed overrules both Suffix and Rime Bias The Suffix Bias in its turn seems to be stronger than the Rime Bias These results suggest that categories with a lower rank in the hierarchy are only effective in case there is no higher-ranked category available There are two possible models that can explain these results First, it is possible that a lower-ranked feature is only active in the selection process when there is no higher-ranked feature available This would mean that features are chosen on the basis of a fall-back strategy For instance, the suffix family is only activated when the left constituent is a novel for-

mation, as in the case of pseudo-words, or when there is no left constituent family available in the mental lexicon. This way of feature selection implies for AML and TiMBL that we need a component that is dynamically tuned to the information in the input. Second, the feature hierarchy can be inclusive, which means that all features affect the selection simultaneously. The highest ranked feature that is available in the input, however, is the one that most effectively determines the selection. There are two considerations that favor the second option. First, recall that including a lower-ranked feature into the feature set on which AML and TiMBL was trained never changed the prediction accuracy of the participants' choices reliably. Second, in 10-fold cross-validation runs over all CELEX compounds, AML and TiMBL reach the highest classification accuracies when the training is based on all three features simultaneously. Further research is required before the question whether the feature hierarchy is inclusive can be solved with certainty.

The third main goal of this study was a comparison of AML and TiMBL with respect to classification accuracy and prediction uncertainty. Comparing the classification and prediction accuracies of AML and TiMBL for existing and novel compounds, we can conclude that, all in all, the models perform equally well. A difference has been found in one case only. Classifying existing compounds taken from CELEX, including the feature Rime in the feature set, led to significantly lower classification accuracies for AML. In all other cases, the observed differences were not reliable, although we should mention that we found a problem of level-interference with TiMBL. When predicting the linkers chosen by participants in the experiment testing the influence of the Rime Bias, including the feature Suffix reduced the prediction accuracy with approximately 10%.

Analyses of the entropy of the choice-distributions of the participants on the one hand and of the models on the other hand revealed that uncertainty is correlated with the strength of the bias in a family. In the case of a neutral bias, both the models and the participants are significantly more uncertain about the appropriate linker than in the case of a strong bias. The relative high uncertainty of participants in the case of a negative bias can be explained by an overall bias for the specific linker for which an experiment is designed, or by a weaker analogical strength of the bias for -ə-. The mean uncertainty of the two models across items in a experimental condition turned out to be identical in all the investigated experiments. We therefore conclude that the models do not differ in their prediction uncertainty.

In this paper, we have focused on the analogical approach to a partly non-deterministic morphological phenomenon. The standard approach to the analy-

sis of morphological phenomena is to formulate formal rules (e.g., Aronoff, 1976; Selkirk, 1982; Lieber, 1981). In these rule-based approaches, the aim is to capture the generalizations that govern the data. Once a formal rule has been formulated on the basis of inspection of the data, the data themselves become irrelevant, because the rule operates independently of the data to its input. Various researchers (e.g., Clahsen, 1999; Marcus, Brinkman, Clahsen, Wiese, & Pinker, 1995; Pinker, 1991; 1997) argue that these symbolic rules have cognitive reality in the brain.

The standard approach has come under attack from connectionist modelers (e.g., Plunkett & Juola, 1999; Seidenberg, 1987; Seidenberg & Hoeffner, 1998; Rueckl, Mikolinski, Raveh, Miner, & Mars, 1997), who exchange symbolic for subsymbolic representations and merge data instances and rules into the connection weights of multi-layered artificial neural networks (ANN). Probably, ANN models will be able to capture the choice of linkers as well. What our simulation results show, however, is that it is not necessary to give up symbolic representations when the goal is to model non-deterministic data. The analogical approach, moreover, is supported by independent psychological evidence that morphological families play a role in language processing (Schreuder & Baayen, 1997; Bertram, Schreuder, & Baayen, 2000; De Jong, Schreuder, & Baayen, 2000), and a sketch of a psycholinguistic spreading-activation model for the selection of linkers can be found in Krott et al. (2001, also chapter 2). We conclude that the analogical approach to morphological rules, in which static symbolic rules abstracted from the data are replaced by dynamic, analogical rules that are linked to and continuously updated by the data, is a fruitful area for future research.

References

Anshen, F. and Aronoff, M.: 1988, Producing morphologically complex words, *Linguistics* **26**, 641–655.

Aronoff, M.: 1976, *Word Formation in Generative Grammar*, MIT Press, Cambridge, Mass.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bertram, R., Schreuder, R. and Baayen, R. H.: 2000, The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Memory, Learning, and Cognition* **26**, 419–511.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 2000, TiMBL: Tilburg Memory Based Learner Reference Guide. Version 3.0, *Technical Report ILK 00-01*, Computational Linguistics Tilburg University.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Fienberg, S.: 1980, *The Analysis of Cross-Classified Categorical Data*, The MIT Press, Cambridge, Mass.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Krott, A., Baayen, R. H. and Schreuder, R.: 2001, Analogy in morphology: modeling the choice of linking morphemes in Dutch, *Linguistics* **39**(1), 51–93.

Krott, A., Schreuder, R., Baayen, R. H. and Dressler, W.: submitted, Analogical effects on linking elements in German compounds.

Lieber, R.: 1981, Morphological conversion within a restrictive theory of the lexicon, *in* M. Moortgat, H. v. d. Hulst and T. Hoekstra (eds), *The Scope of Lexical Rules*, Foris, Dordrecht, pp. 161–200.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R. and Pinker, S.: 1995, German inflection: The exception that proves the rule, *Cognitive Psychology* **29**, 189–256.

Mattens, W. H. M.: 1984, De voorspelbaarheid van tussenklanken in nomi-

nale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343.

Pinker, S.: 1991, Rules of language, *Science* **153**, 530–535.

Pinker, S.: 1997, Words and rules in the human brain, *Nature* **387**, 547–548.

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Rietveld, T. and Van Hout, R.: 1993, *Statistical Techniques for the Study of Language and Language Behaviour*, Mouton de Gruyter, Berlin.

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S. and Mars, F.: 1997, Morphological priming, fragment completion, and connectionist networks, *Journal of Memory and Language* **36**(3), 382–405.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Seidenberg, M.: 1987, Sublexical structures in visual word recognition: Access units or orthographic redundancy, *in* M. Coltheart (ed.), *Attention and Performance XII*, Lawrence Erlbaum Associates, Hove, pp. 245–263.

Seidenberg, M. and Hoeffner, J.: 1998, Evaluating behavioral and neuroimaging data on past tense processing, *Language* **74**, 104–122.

Selkirk, E.: 1982, *The Syntax of Words*, The MIT Press, Cambridge.

Skousen, R.: 1989, *Analogical Modeling of Language*, Kluwer, Dordrecht.

Van den Toorn, M. C.: 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is (Linking phonemes in compounds with derived forms as first constituents), *De Nieuwe Taalgids* **74**, 197–205.

Van den Toorn, M. C.: 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is (Linking phonemes in compounds with loanwords as first constituents), *De Nieuwe Taalgids* **74**, 547–552.

Van den Toorn, M. C.: 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I)), *De Nieuwe Taalgids* **75**(1), 24–33.

Van den Toorn, M. C.: 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for the regulation of linking phonemes in nominal compounds II), *De Nieuwe Taalgids* **75**(2), 153–160.

Vance, T. J.: 1980, The psychological status of a constraint on Japanese consonant alternations, *Linguistics* **18**, 145–167.

Appendices

Appendix A

Materials of Experiment 1: left constituent and right constituent (number of responses with a linker, number of responses without a linker). Capital letters mark word stress.

L1: Positive Rime Bias:
LANtan organisatie (16, 4); VAneid kooi (18, 2); PEUzeid steun (18, 2); KApeid gedrag (17, 3); HOran oord (16, 4); MOEveid voer (18, 2); NOgan plicht (19, 1); GOEran probleem (16, 4); VEEpleid milieu (15, 5); BAlan geschiedenis (15, 5); PLAveid paar (19, 1); LUIsan pensioen (18, 2); MIJstan commissie (18, 2); BOElan niveau (15, 5); KOlan controle (12, 8); DAkeid republiek (16, 4); LUchan conflict (15, 5); BOEneid stam (14, 6); TOpleid gezicht (17, 3); ZApleid verzameling (16, 4); KEEzeid waarde (17, 3); GROtan aanbod (14, 6); VIJzan dienaar (17, 3); POEkeid hok (18, 2)

L2: Neutral Rime Bias:
Oloe corps (12, 8); MARvoe verzekering (13, 7); TOtroe galerij (8, 12); BOdap regeling (16, 4); KIJdap structuur (13, 7); VEUnen pensioen (9, 11); STIEvap karakter (15, 5); DROlen oord (16, 4); TAzoe tak (13, 7); PAgoe toestand (13, 7); BLOstoe hut (8, 12); MIEfap element (14, 6); SCHIJlen middel (10, 10); PLOElen element (10, 10); BIEvap zone (15, 5); BOEdap middel (19, 1); VILnoe vlees (11, 9); POEnen organisatie (6, 14); KRAzen conflict (10, 10); POERgoe vrouw (11, 9); KOdap beleid (14, 6); ZOzen zone (9, 11); PUIbap rust (19, 1); DULLen rust (12, 8)

L2: Negative Rime Bias:
NApla bond (7, 13); TUIzem dienaar (15, 5); BIEzel waarde (12, 8); SILda tong (13, 7); KLAvij structuur (9, 11); DRAsij regeling (12, 8); WONkel geschiedenis (11, 9); BRAnij hulp (13, 7); TIKsem aanbod (14, 6); BISSel probleem (12, 8); PLUIvij karakter (12, 8); PLOdem plicht (16, 4); POEkrij conferentie (10, 10); ARta vel (11, 9); STIJza kas (10, 10); LIEsem niveau (18, 2); TISSel milieu (6, 14); DUISkra zee (7, 13); STAlem controle (8, 12); DRUImel gedrag (9, 11); SOERkwa kop (10, 10); VOEnij beleid (12, 8); VAjel gezicht (15, 5); KROEsem commissie (12, 8)

Appendix B

Materials of Experiment 2 left constituent and right constituent (number of responses according to the constituent, number of responses according to the suffix) vluchteling gezicht (20, 0), voorziening regeling (11, 9), belasting kas (13, 7), vreemdeling republiek (20, 0), tiener gedrag (18, 2), tweeling kop (14, 6), kaper hulp (3, 17), woning kooi (17, 3), zuigeling probleem (19, 1), luidspreker hok (7, 13), leerling vel (20, 0), klapper galerij (14, 6), veiling commissie (9, 11), waterleiding aanbod (15, 5)

Materials of Experiment 3 left constituent and right constituent (number of responses according to the constituent, number of responses according to the rime) handel geschiedenis (20, 0), idee waarde (13, 7), bij controle (13, 7), ezel tong (17, 3), levensmiddel organisatie (19, 1), specerij zee (20, 0), dominee pensioen (16, 4), engel dienaar (19, 1), schilderij paar (20, 0), duivel plicht (20, 0)

Appendix C

Materials of Experiment 4 List A definition plus left and right compound constituent (number of responses according to the bias of the suffix, number of responses according to the bias of the letter combination, number of other responses)

Een 'PLOEver' is een boomsoort	PLOEver gried (12,5,3)
In een glas 'WILter' zit veel alcohol	WILter_boest (11,8,1)
Een 'VIEber' is een blaasinstrument	VIEber_gedij (5,12,3)
Een 'VOEStegaar' is een verdedigingstactiek	VOEStegaar_sien (9,7,4)
Een 'MOEnaar' is een visvergunning	MOEnaar gezoel (9,2,9)
Iets wat zeldzaam is noemen we een 'BOEzaar'	BOEzaar turei (13,4,3)
Mediterrane vegetatie heet ook wel 'ROEzing'	ROEzing nast (12,2,6)
'PRIEling' is een kruidensoort	PRIEling faren (14,3,3)
Een 'KRONving' is een muziekstuk	KRONving doef (11,1,8)
'BinTIST' is een Oosters gerecht	binTIST_zaste (16,3,1)
Een 'baraFIST' is een opslagtank	baraFIST modee (13,5,2)
'GisoFIST' is een Belgisch biermerk	gisoFIST buroop (13,7,0)

Materials of Experiment 4. List B definition plus left and right compound constituent (number of responses according to the bias of the suffix, number of responses according to the bias of the letter combination, number of other responses)

Iemand die graag 'ploeft' is een 'PLOEver'	PLOEver_gried (17,3,0)
Iemand die 'wilt' is een 'WILter'	WILter_boest (14,5,1)
Een persoon die goed 'viebt' is een 'VIEber'.	VIEber_gedij (13,6,1)
Iemand die graag 'voest' is een 'VOEStegaar.	VOEStegaar_sien (10,5,5)
Degene die 'moent' is de 'MOEnaar'.	MOEnaar_gezoel (16,2,2)
De persoon die 'boest' is de 'BOEzaar'	BOEzaar_turei (9,4,7)
Het 'roezen' van iets heet de 'ROEzing'.	ROEzing_nast (12,3,5)
Het 'prielen' van iets is de 'PRIEling'.	PRIEling_faren (15,2,3)
Het resultaat van het 'kronven' is de 'KRONving'	KRONving_doef (12,0,8)
Iemand die een 'bint' bespeelt is de 'binTIST'	binTIST_zaste (17,3,0)
De 'baraaf' wordt bespeeld door de 'baraFIST'.	baraFIST modee (18,2,0)
De 'gisoof' wordt gemaakt door de 'gisoFIST'	gisoFIST buroop (19,1,0)

Appendix D

Rules applied to the CELEX compounds with a token frequency of at least one in a corpus of 42 million (22,966 compounds): If first constituent

- ends in schwa plus sonorant, use *-∅-*.
- ends in a full vowel, use *-∅-*.
- has the feature <+human> and ends in *-er*, *-eur*, *-ier*, *-aar*, or *-air*, use *-s-*.
- has the feature <+human> and ends in *-ist*, *-erik*, *-es*, *-in*, *-aan/-iaan*, *-ling/-eling*, *-uur*, *-ant*, *-ent*, *-aat*, *-iet*, *-aal*, *-eel*, *-iel*, *-loog*, or *-graaf*, use *-e(n)-*.
- has the feature <+human> and ends in *-ette*, use *-∅-*.
- has the feature <+human> and ends in *-or*, use *-s-* or *-e(n)-*.
- has the feature <-human> and ends in *-uur*, in *-ant*, in *-iet*, in *-aal*, in *-eel*, in *-iel*, in *-loog*, in *-graaf*, *-air*, or *-or*, use *-∅-*.
- has the feature <-animate> and ends in *-er*, *-eur*, *-ier*, *-ette*, or *-in*, use *-∅-*.
- has the feature <-animate> and ends in *-er*, *-eur*, *-ier*, *-ette*, or *-in*, use *-∅-*.
- has the feature <-countable> and ends in *-teit/-iteit*, *-schap*, *-ing*, or *-dom*, use *-s-*.
- has the feature <+countable> and ends in *-teit/-iteit*, *-schap*, *-dom*, *-dij/ -erij/-arij*, or *-nis*, use *-e(n)-*.
- has the feature <-countable> and ends in *-isme*, *-nis*, *-ij/-erij/-arij*, or *-ade/ -ide/-ode*, use *-s-*.
- has the feature <+countable> and ends in *-isme*, *-nis*, or *-ade/-ide/-ode*, use *-n-*.
- has the feature <+countable> and ends in *-ing*, use *-∅-*.
- ends in *-heid*, or *-(t)je*, use *-s-*.
- ends in *-te/-de*, *-sel*, *-sie/-tie*, *-um*, *-theek*, *-aris*, or *-us*, use *-∅-*.

Semantic Effects

This chapter will be published as Andrea Krott, Loes Krebbers, Robert Schreuder, and R. Harald Baayen: Semantic influence on linkers in Dutch noun-noun compounds, *Folia Linguistica*.

Abstract

As in many other languages, the constituents of nominal compounds in Dutch are often separated by a linking element. This study investigates to what extent form and semantic properties of the right constituents in Dutch compounds affect the choice of the linker. Using both lexical statistics and experimentation, we show that the left and right constituent families affect the choice of the linker independently of the semantic categories of the left and right constituents themselves. We also show that the choice of the linker is co-determined by the animacy and concreteness of the left constituent. No role for the semantic class of the head constituent was observed in the experiment. Apparently, linkers are non-canonical suffixes in the sense that their occurrence is codetermined by the form properties of the constituent to their right.

Introduction

In many languages, elements known as connectives, interfixes, linking morphemes, and linkers, may occur between the two constituents of compounds. Sometimes, the occurrence of such linkers can be predicted on phonological grounds as in Zoque, a Mixe-Zoquean language spoken in Mexico. Zoque has a nominal compound formation in which the linking element is a vowel that is identical to the vowel in the preceding syllable (Herrera, 1995). In Germanic languages such as German and Dutch, the principles governing their distribution are less clear. The distribution of linkers in German appears to be governed by a complex set of factors (see, e.g., Dressler, Libben, Stark, Pons, & Jarema, 2001; Fuhrhop, 1998). Although Dutch is closely related to German, the linkers in Dutch have different properties, possibly because, in contrast to German, modern Dutch no longer has productive case morphology. Historically, the Dutch linkers can be traced to the case endings that existed in medieval Dutch. However, the original functionality of the linking elements as case suffixes is absent in modern Dutch.

More than a third of Dutch noun-noun compounds contain a linker connecting the two main constituents.[1] Usually *-s-* or one of the orthographic variants *-en-* and *-e-* appear as a linker (e.g., *bevolking+s+getal* 'number of population', *boek+en+kast* 'bookcase', *zon+e+schijn* 'sunshine').[2] The usage of these linkers reveals considerable variation and unpredictability. Existing rule-based descriptions report various morphological, phonological, and semantic factors which seem to govern their choice (e.g., Van den Toorn, 1981a; 1981b; 1982a; 1982b; Mattens, 1984; Haeseryn, Romijn, Geerts, De Rooij, & Van den Toorn, 1997; Booij & Van Santen, 1995). However, almost every rule comes with a large number of exceptions. Taking all phonological and morphological rules together that are described in the literature[3], one can apply them to only 51% of all CELEX compounds and correctly predict only 63% of this subset. We therefore may conclude that a rule-based account for Dutch linkers is observationally inadequate.

Krott, Baayen, & Schreuder (2001, also chapter 2) and Krott, Schreuder, & Baayen (in press, also chapter 3) argue that the choice of linkers in Dutch is governed by

[1] Of all noun-noun compounds listed in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) 35% are formed with a linker

[2] A description of spelling variants *-en-* and *-e-* can be found, e.g., in the *Woordenlijst Nederlandse Taal* (1995)

[3] For a complete list of phonological and morphological rules, see Appendix D of Krott, Schreuder, & Baayen (in press, also chapter 3). Semantic rules were not taken into account because semantic information in CELEX is not available.

analogy. Using an off-line cloze task in which participants had to form novel compounds from two Dutch nouns, they show that the usage of linkers in novel compounds can be predicted with a high degree of accuracy on the basis of analogy to the forms of existing compounds sharing the left or the right constituent of a given target compound, for instance, *schaap-?-oog*, 'sheep-eye'. We refer to the set of compounds sharing the left constituent (*schaap* 'sheep' in this example) as the Left Constituent Family (*schaap+en+bout*, 'leg of mutton', *schaap+s+kooi*, 'sheep fold', *schaap+herder*, 'shepherd', etc.), and we refer to the set of compounds sharing the right constituent (*oog* 'eye' in the present example) as the Right Constituent Family (*uil+e+oog* 'owl's eye', *spleet+oog* 'slant eye', *glas+oog* 'glass eye', etc.). One can predict the choice of the linker for a novel compound on the basis of the distribution of linkers in its Left and Right Constituent Families. For instance, if *schaap* occurs as a left constituent mostly in compounds containing the linking *-en-* (70% in CELEX), there is a high chance that a novel compound with *schaap* as the left constituent would also be built with *-en-*.

The strongest analogical factor predicting linkers appears to be the bias of the Left Constituent Family. In addition to the Constituent Family, experiments with pseudo-stems followed by existing suffixes as left constituents showed effects of the bias of the suffix and the rime of the left constituent. The bias of the suffix appears to be the second strongest factor overruling the bias of the rime. Apart from the effects of the left constituent, there is also evidence for a smaller, but statistically reliable effect of the bias of the Right Constituent Family.[4] Explicit computational models for analogy (AML: Skousen, 1989; TiMBL: Daelemas, Zavrel, Van der Sloot, & Van den Bosch, 2000) provide excellent fits to the empirical data as well as to the distributional patterns in CELEX.

The analogical form effect of the Right Constituent Family on the choice of the linker is surprising as the left constituent is usually taken to be the prime determiner (see, e.g., Booij, 1996; Mattens, 1970). First, etymologically, both *-en-* and *-s-* developed out of inflectional suffixes, i.e. markers for genitive singular or nominative plural. The linker *-en-* is still restricted to first constituents that select the suffix *-en* for the formation of noun plurals. Second, there is experimental evidence that adding

[4]The rules in the literature focus on the properties of the left constituent and never consider the right constituent as a possible factor. Note that the effect of the Right Constituent Family cannot be accounted for by means of rules that would be sensitive to the phonological or morphological properties of the right constituent. A statistical survey of 22966 Dutch compounds shows that the onset and, if present, the prefix of the second constituent can be used to predict only 64.5% of the linkers, which is identical to the percentage of compounds with no linker (the default) and thus could be attained by always chosing the linker with the a-priori maximum likelihood.

a linker to a first constituent may activate plural semantics (Schreuder, Neijt, Van der Weide, & Baayen, 1998). Third, phonologically, linkers belong to the first constituents of compounds. The linker always groups with the final syllable of the first constituent (e.g., *koning+s+kind* 'king's child'), even when the second constituent is separated from the first in contractions such as *varken+s- en schap+e+vlees* 'pork and mutton'. Finally, left constituents sometimes undergo vowel alternation in combination with a linker (compare *schip+breuk*, 'shipwreck', *scheep+s+werf*, 'shipyard'), suggesting that the left constituents and their linkers might also be interpreted as allomorphs. Considered jointly, these observations strongly suggest that the linker groups with the left constituent and that compounds with linkers are left-branching structures.

The strong analogical force of the Left Constituent Family reported by Krott et al. (2001, also chapter 2) is in line with the above considerations, while the weaker but statistically reliable analogical force of the Right Constituent Family that they report is surprising and requires further investigation. The aim of the present paper is to explore whether the observed analogical effect of the Right Constituent Family might be not an analogical effect based on the pure forms of the Right Constituent Families, but rather an analogical effect based on the semantic properties of the Right Constituents. Thus, we focus on the question whether it is the set of compounds sharing the Right Constituent with the target compound that forms the analogical basis for the choice of the linker or whether it is the set of compounds sharing the semantic class of the Right Constituent with the target compound that forms the analogical basis. Returning to the example of *schaap-?-oog*, the question is whether we should consider the set of compounds having *oog* as right constituent, or whether we should consider the set of compounds that have, for instance, a concrete noun as right constituent.

Van den Toorn (1982a) mentions several semantic factors that might be relevant. These factors fall into two types. First, the semantic class of a constituent might play a role. First constituents that are mass nouns, for instance, seem to occur predominantly without a linker (e.g., *papier+handel* 'paper trade'), though this is not always the case (*tabak+s+rook* 'tabacco smoke'). Second, the semantic relation between the constituents seems to have an influence on the choice of the linker. For example, if the first constituent is the logical object of the second constituent, the constituents tend to be connected without a linker (e.g., *boek+verkoper* 'book seller', but again there are many exceptions, e.g., *gezin+s+planning* 'family planning'). We will restrict our focus to the first kind of semantic factors, the semantic

class of the constituents.

Some preliminary evidence for an effect of the semantic class of the constituents has already been found in post-hoc simulation studies in which responses of participants have been modeled with TiMBL. The responses were produced in two cloze tasks which orthogonally varied the bias of the Left and Right Constituent Family (Krott et al., 2001, also chapter 2). Simulation studies with TiMBL revealed optimal prediction accuracies when the analogy was based not only on the left constituent, i.e. the Left Constituent Family, but also on information concerning the semantic class of the right constituent. Prediction accuracy did not improve any further by additionally taking the Right Constituent Family into account. These results suggest that the form effect of the Right Constituent Family might indeed be a semantic effect. However, these post-hoc analyses are inconclusive by themselves and require supplementation by an independent factorial experiment explicitly addressing the potential role of semantic categories.

In what follows, we first present some lexical statistics concerning the relation between the use of linkers in Dutch compounds and the semantic class of the left and right constituents. Next, we discuss a factorial experiment designed to clarify the potential effect of semantic features on the choice of linkers in novel compounds, following which we reanalyze the experiments reported in Krott et al. (2001, also chapter 2) with respect to the role of the semantics of the left and right constituents.

Lexical statistics

In order to ascertain the effect of the semantics of both left and right constituents, we investigated the 6949 compounds in the families of the first two experiments reported in Krott et al. (2001, also chapter 2). For these compounds, we have annotated the left and right constituents with the following semantic categories: abstract versus concrete, and animate versus inanimate. Within the category of animate nouns, we distinguished between human versus animal, and within the category of inanimate we distinguished between plant versus other. Table 4.1 gives an overview over the distribution of linkers across these 6949 compounds when we partition these compounds according to the semantics of the first and second constituents. A partition into abstract and concrete first constituents reveals, for instance, that *-en-* prefers concrete first constituents. An independent partitioning according to the animacy of the first constituent shows that animate first constituents prefer *-en-* and that no linker is preferred for inanimate nouns. Partitionings according to the

Table 4.1: Numbers of linking possibilities for different semantic classes of the left and right constituents of 6949 Dutch compounds.

Constituent	Semantic Class	*-s-*	*-en-*	*-∅-*
first				
	abstract	1801	172	1471
	concrete	559	1007	1939
	animate	274	510	195
	inanimate	2086	669	3215
second				
	abstract	1818	415	1497
	concrete	542	764	1913
	animate	157	79	345
	inanimate	2203	1100	3165

second constituents show different distributions.

A more informative way of summarizing the distribution of the linkers as a function of semantic categories is to construct a classification tree using a non-parametric technique, CART (Breiman, Friedman, Olshen, & Stone, 1984). CART is useful for classification problems with one or more predictor variables (here: the semantic class) and one response variable (here: the linker). The statistical model is fitted by binary recursive partitioning of the data, which means that the dataset is successively split up into increasingly homogeneous subsets with different values of the predictor variable (different semantic classes). Each split partitions the data into two subsets while maximizing the difference in the relative proportions of linkers. This process results in a classification tree.

Model selection in CART analyses is accomplished by means of cost-complexity pruning, a technique for finding the smallest (most parsimoneous) tree with low heterogeneity of the leaves. The left panel of Figure 4.1 plots the cross-validation score function. The horizontal axis plots the size of the classification tree, the vertical axis plots the corresponding deviance (calculated using 10-fold cross-validation). The deviance is a measure of average node heterogeneity. The upper axis shows the mapping between tree size and the cost-complexity parameter α (by increasing α, the size of the tree is penalized more heavily). We chose a quite conservative α of .0145, following the advice of Breiman et al. (1984). The resulting pruned tree is shown in the right panel of Figure 4.1. Table 4.2 lists the percentages of linkers for the leaves of the pruned tree as it is presented in Figure 4.1. The length of the vertical lines represents the amount of deviance accounted for by a partic-

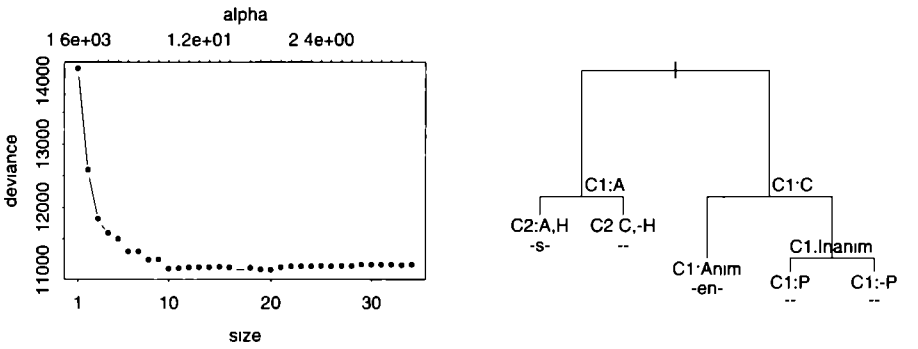


Figure 4.1: CART analysis of the semantic classes of the constituents of 6949 Dutch compounds as predictor variable and linker (*-en-*, *-s-*, and – in the case of a zero realization) as the response variable; left panel: plot of deviance versus tree size for sequences of subtrees; right panel: pruned classification tree; C1 = first constituent; C2 = second constituent; A = abstract; C = concrete; H = human being; Anim = animate; Inanim = inanimate; P = plant.

Table 4.2: Percentages of linkers for the leaves of the pruned tree of Figure 4.1 (see the legend of Figure 4.1 for further details of notation).

Node	*-en-* (%)	*-∅-* (%)	*-s-* (%)
C1:A;C2:A,H	4	36	60
C1:A;C2:C,-H	6	65	29
C1:C,Anim	52	20	28
C1:C,Inanim,P	44	56	1
C1:C,Inanim,-P	16	74	10

ular split. The largest deviance and therefore the largest predictive power is given by the partition into abstract and concrete first constituents. The next highest deviance is reached by the split into first animate and inanimate constituents. The latter are further divided into plants and non-plants. The semantic class of the second constituent seems to be less relevant. The only predictive split appears to be the division into abstract nouns and human beings on the one side and concrete objects that are not human beings on the other side. Concreteness and animacy of the first (left) constituent emerge as strong predictors of the linkers in our data. For right constituents, it seems to matter to some extent whether they are abstract or concrete and whether they are human beings.

Summing up, the concreteness and animacy of the first constituent emerge from

this analysis as reliable predictors of the linkers The predictive force of the concreteness of the second constituent is weak The next section addresses the question whether it is still strong enough to guide the decisions of participants in a cloze task

A production experiment

Method

Materials We constructed three sets of left constituents (L1, L2, L3) and four sets of right constituents (R1, R2, R3, R4) Each set contained 10 Dutch nouns Given the results of the CART analysis, we considered animacy and concreteness as the main important semantic features and the feature 'human-being' as an additional feature potentially important for right constituents Therefore, we chose the following experimental sets The groups of left constituents contained abstract (L1), concrete-inanimate (L2), and concrete-animate nouns (L3) The sets of right constituents contained abstract (R1), concrete-inanimate (R2), concrete-human (R3) and concrete-animal nouns (R4) We made sure that all left constituents can be combined with the linker *-en-* In addition, all constituents have a bias against being combined with a linker, i e at least 60% of all compounds in the Constituent Families occur without a linker (L1 mean 82 7%, range 64%-97%, L2 mean 81 4%, range 71 4%-100%, L3 mean 81 0%, range 63 6%-100%, R1 mean 75 6%, range 61 5%-100%, R2 mean 83 8%, range 60 0%-100%, R3 mean 90 7%, range 66 7%-100%, R4 mean 92 6%, range 60 0%-100%)

Each of the three sets of left constituents (L1, L2, L3) was combined with the four sets of right constituents (R1, R2, R3, R4) to form pairs of constituents for new compounds in a factorial design with two factors Semantic Class of the Left Constituent (abstract, concrete-inanimate, concrete-animate) and Semantic Class of the Right Constituent (abstract, concrete-inanimate, concrete-human, concrete-animal) None of these compounds is attested in the CELEX lexical database with a token frequency higher than zero All have a high degree of semantic interpretability The Appendix lists all experimental items The 3 x 4 x 10 = 120 experimental items were divided over three lists List 1 contained the compounds of the factorial combinations L1-R1, L2-R4, and L3-R3 List 2 contained the compounds of the combinations L1-R3, L2-R2 and L3-R4 List 3 contained the compounds of the combinations L1-R4, L2-R1, and L3-R2, and List 4 contained the compounds of the combinations L1-R2, L2-R3, and L3-R1 In this way, each participant saw a given

constituent only once. We constructed a separate randomized list of the 3 x 10 = 30 pairs of compound constituents for each participant.

Procedure. The participants performed a cloze-task. An experimental list of items was presented to the participants in written form. Each line presented two compound constituents separated by two underscores. We asked the participants to combine these constituents into new compounds and to specify the most appropriate linker, if any, at the position of the underscores, using their first intuitions. Occasionally, the first constituent may change its form when it is combined with a linker (e.g., *schip* ('ship') appears as *scheep* in the compound *scheepswerf* ('shipyard')). The instructions clarified that these changes were not of interest and could be ignored. We told the participants that they were free to use *-en-* or *-e-* as spelling variants of the linker *-en-*. The experiment lasted approximately 10 minutes.

Participants. Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch. The participants were divided into three groups, one for each experimental list.

Results and discussion

One participant produced unexpected, non-standard letter sequences for three stimuli. These responses were classified as errors and excluded from the statistical analyses. Table 4.3 summarizes the percentages of the responses for the twelve experimental conditions. The Appendix lists the individual words together with the counts of the responses.

The counts of *-s-*, *-en-*, and *-∅-* responses for a given word are not independent — they always sum up to 20, the total number of participants. In order to bring the data in line with the requirements of standard multivariate methods, we divided the number of *-en-* and *-s-* responses by the number of *-∅-* responses. A multivariate analysis of variance of the logarithms of the resulting ratios[5] revealed a main effect of the Semantic Class of the Left Constituent, but no effect of the Semantic Class of the Right Constituent, and no interaction of both factors (left semantic class: $F_2(2,108) = 16.8$, $p < .001$; right semantic class: $F_2(3,108) = 1.1$, $p = .374$).

The way in which the Semantic Class of the left Constituent affects the responses of the participants is summarized in Figure 4.2. Responses with the linking *-s-* (solid line) occur predominantly with abstract left constituents. By contrast, *-en-* responses (dotted line) are least frequent with abstract constituents, but common for concrete, and even more common for animate concrete left constituents. Re-

[5]Counts equal to zero were set to 0.1 before taking the logarithm.

Table 4.3: Percentages and numbers of selected linkers when varying the Semantic Class of the Left and Right Constituent. 'c-': concrete.

		Right Constituent							
Left		abstract		c-inanimate		c-human		c-animal	
Constituent		%	#	%	#	%	#	%	#
abstract									
	en	28.5	57	34.0	68	31.0	62	30.0	60
	s	29.5	59	23.0	46	19.5	39	25.5	51
	∅	42.0	84	43.0	86	48.5	97	44.5	89
inanimate									
	en	56.0	112	43.5	87	55.5	111	44.5	89
	s	4.5	9	4.5	9	1.5	3	9.0	18
	∅	39.5	79	52.0	104	43.0	86	46.6	93
animate									
	en	75.0	150	77.5	155	54.0	108	52.0	104
	s	2.0	4	2.5	5	5.5	11	5.5	11
	∅	23.0	46	20.0	40	40.5	81	42.0	84

sponses with -∅- (dashed line) are slightly less common for concrete animate left constituents. This pattern of results is quite similar to the general pattern in the Dutch lexicon as summarized in Table 4.2 above.

Do the Left and Right Constituent biases co-determine the choice of the linker in addition to the semantic class of the Left Constituent? More specifically, does the absence of a semantic effect for the Right Constituent imply that no effect of the bias of the Right Constituent should be observable? A post-hoc multivariate analysis of covariance revealed reliable effects of both the Left and Right Constituent Bias in addition to the factorially established effect of the Left Semantic Class (Left Constituent Family: $F_2(2,109) = 6.5$, $p < .001$; Right Constituent Family: $F_2(2,109) = 2.5$, $p = .047$; Left Semantic Class: $F_2(2,109) = 18.6$, $p < .001$). We also observed an interaction of the Semantic Class of the Left Constituent and the Bias of the Left Constituent Family ($F_2(4,109) = 2.6$, $p = .009$). We conclude that, apparently, the Right Constituent Family is a factor in its own right that cannot be reduced to a semantic effect of the right constituent.

Recall that the CART analysis of the relation between the semantic categories and the linkers revealed a weak but reliable effect for the concreteness of the right constituent (Figure 4.1, Table 4.2). The present experimental results suggest that the variability in the lexicon is too large to allow individual language users to make

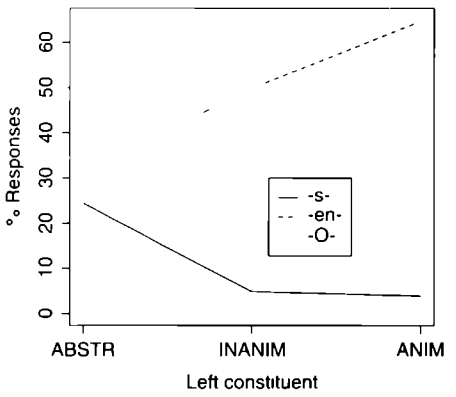


Figure 4.2: Percentages of *-en-*, *-s-* and *-∅-* responses for different Semantic Classes of the left constituent (ABSTR: abstract; INANIM: inanimate; ANIM: animate).

use of the observed distributional pattern. It is possible that the present experimental paradigm is not sensitive enough to register potential semantic effects of the right constituent. However, given that it is sensitive enough to reveal a reliable effect for the Right Constituent bias and a clear effect of the semantics of the Left Constituent, we have to conclude that, at the very least, the effect of the Right Constituent bias is much stronger than the potential effect of the semantics of the Right Constituent.

These results raise the question whether the semantic effect reported by Krott et al. (2001, also chapter 2) on the basis of two cloze tasks mentioned in the introduction is reliable. A post-hoc logit analysis of the EN-experiment with semantic class as covariate revealed main effects for the Left and Right Constituent families (Left Constituent Family: $F_2(2,141) = 92.18$, $p < .001$; Right Constituent Family: $F_2(2,141) = 11.68$, $p < .001$) as well as a main effect of the Semantic Class of the Left Constituent ($F_2(5,164) = 5.70$, $p < .001$). No such effect could be observed for the right constituent ($F_2(5,164) < 1$). As in the present experiment, a similar interaction between the Semantic Class of the left constituent and the Bias of the Left Constituent Family was visible ($F_2(8,141) = 2.22$, $p = .029$). Analyses of the S-experiment revealed the same pattern of results.[6] These post-hoc analyses

[6]Semantic Class of the Left Constituent: $F_2(5,173) = 3.78$, $p = .003$); Semantic Class of the Right Constituent: $F_2(4,173) < 1$; Left Constituent Family: $F_2(2,153) = 124.65$, $p < .001$; Right Constituent

parallel the results obtained in the present experiment and confirm that the Right Constituent Family bias cannot be reduced to a semantic effect. Apparently, the slight increase in prediction accuracy reported by Krott et al. (2001, also chapter 2) that they obtained using TiMBL is not robust, and, in fact, inclusion of the semantic information for the second constituent does not lead to a statistically significant improvement in performance in their experiments (EN-experiment: 86.6% versus 79.9%, $\chi^2_{(1)} = 2.74$, $p = .0977$; S-experiment: 88.4% versus 87.3%, $\chi^2_{(1)} = .02$, $p = .875$).

General discussion

This study addressed the question whether the Right Constituent Family affects the choice of linkers in Dutch noun-noun compounds, an analogical effect across complex words sharing constituents, or whether the semantic category of the right constituent is the crucial factor at issue. A statistical survey of 6949 Dutch compounds and the semantic categories of their constituents revealed that the concreteness or abstractness of the right constituent is a minor predictor of the linker compared to the semantic class of the left constituent. However, a factorial experiment using a cloze task revealed a reliable effect of the Left Semantic Class, but no effect whatsoever of the Right Semantic Class. A post-hoc analysis revealed clear effects of both the Left and Right Constituent Families and an Interaction of the Left Semantic Class and the Left Constituent Family. Re-analyses of the experiments reported by Krott et al. (2001, also chapter 2) yielded the same pattern of results.

The failure to find any influence of the Right Semantic Class in combination with the clearly observable robust effect of the Right Constituent Family falsifies our initial hypothesis that the effect of the Bias of the Right Constituent Family might in fact be an effect of the semantic class of the right constituent. We have to conclude that the choice of the linker in Dutch is analogically co-determined by the distribution of linkers in the set of compounds sharing the right constituent.

What then, is the morphological status of the linkers in Dutch? Clearly, Dutch linkers are not normal suffixes. Whether or not a suffix can be attached to a base word may depend on the phonological, morphological, and semantic properties of the base. But, to our knowledge, normal suffixes never depend on the properties of what follows to their right.

Family: $F_2(2,153) = 9.34$, $p < .001$; Interaction between the Semantic Class of the Left Constituent and the Bias of the Left Constituent Family ($F_2(4,153) = 9.64$, $p < .001$.

Although, as mentioned in the introduction, linkers resemble normal suffixes in their strong etymological, semantic, and phonological dependence on the left constituent, there is also evidence that they may not form very strong units with their left constituents. For instance, Kehayia, Jarema, Tsapkini, Perlak, Ralli, & Kadzielawa (1999) report that left constituents followed by linkers in Polish and Greek compounds are effective primes only when their combination occurs as a separate (inflected) word in the language. Without such support, left constituents followed by linkers do not prime, which is not what one would expect if the linker and the left constituent would form a unit at some level of representation in the mental lexicon.

The unexpected role for the right constituent on the choice of the linker in Dutch may be due to the absence of a clear functional role for linkers in this language. From a historical perspective, the following sequence of events may have occurred. Initially, various nominal case endings occurred in compounds. Many such compounds, especially those enjoying a frequent use, were probably stored in the mental lexicon (Van Jaarsveld & Rattink, 1988). Following the loss of the nominal case system, the only place where nominal case endings were retained in great numbers was nominal compounding, where they persisted thanks to their being stored in the mental lexicon. In the absence of a clear functional role, each new generation of language learners is faced with the problem of having to use the standard forms as in current use in the community without having recourse to a clear-cut systematicity for predicting the correct form for existing words and for the formation of new compounds. In such a situation, all possible sources of information might be useful. One such source of information might be the semantic classes of the constituents. In modern Dutch, the abstractness versus the concreteness of the modifying constituent might be a growing source of systematicity for a functional re-interpretation of the linkers from a case-marker to a marker of semantic class. But we suspect that as long as such a process of re-interpretation has not been fully completed, all available information, including the distributional information contained in the Right Constituent Family, is used to optimize the chances of the learner to conform to the current norms in the society.

Note, finally, that there are two ways in which our data on the analogical nature of the choice of linkers in Dutch can be interpreted. On the one hand, it may be argued that this kind of analogical word formation is typical for language domains that have become more or less chaotic due to historical change. On the other hand, it may be that analogy is much more pervasive and underlies phenomena traditionally analyzed as rule-governed. From this second perspective, the Dutch linkers provide

an excellent window on the general properties of analogy. Future research will have to clarify the merits of these contrasting views.

References

Baayen, R H , Piepenbrock, R and Gulikers, L 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA

Booij, G and Van Santen, A 1995, *Morfologie De Woordstructuur van het Nederlands* (Morphology The Structure of Dutch Words), Amsterdam University Press, Amsterdam

Booij, G E 1996, Verbindingsklanken in samenstellingen en de nieuwe spellingregeling (linking phonemes in compounds and the new spelling system), *Nederlandse Taalkunde* **2**, 126–134

Breiman, L , Friedman, J , Olshen, R and Stone, C 1984, *Classification and Regression Trees*, Wadsworth International Group, Belmont, California

Daelemans, W , Zavrel, J , Van der Sloot, K and Van den Bosch, A 2000, TiMBL Tilburg Memory Based Learner Reference Guide Version 3 0, *Technical Report ILK 00-01*, Computational Linguistics Tilburg University

Dressler, W U , Libben, G , Stark, J , Pons, C and Jarema, G 2001, The processing of interfixed German compounds, *in* G E Booij and J Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp 185–220

Fuhrhop, N 1998, *Grenzfalle Morphologischer Einheiten (Border Cases of Morphological Units)*, Stauffenburg, Tuebingen

Haeseryn, W , Romijn, K , Geerts, G , de Rooij, J and van den Toorn, M 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen

Herrera, Z E 1995, *Palabras Estratos y Representaciones Temas de Fonologia Lexica en Zoque*, El Colegio de Mexico

Kehayia, E , Jarema, G , Tsapkini, K , Perlak, D , Ralli, A and Kadzielawa, D 1999, The role of morphological structure in the processing of compounds The interface between linguistics and psycholinguistics, *Brain and Language* **68**, 370–377

Krott, A , Baayen, R H and Schreuder, R 2001, Analogy in morphology modeling the choice of linking morphemes in Dutch, *Linguistics* **39**(1), 51–93

Krott, A , Schreuder, R and Baayen, R H in press, Analogical hierarchy exemplar-based modeling of linkers in Dutch noun-noun compounds, *in* R Skousen (ed), *Analogical Modeling An Exemplar-Based Approach to Language*

Mattens, W H M 1970, *De indifferentialis Een onderzoek naar het anumerieke gebruik van het substantief in het algemeen bruikbaar Nederlands*, Van Gorgum, Assen

Mattens, W H M 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343

Schreuder, R , Neijt, A , Van der Weide, F and Baayen, R H 1998, Regular plurals in Dutch compounds linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573

Skousen, R 1989, *Analogical Modeling of Language*, Kluwer, Dordrecht

Van den Toorn, M C 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is (Linking phonemes in compounds with derived forms as first constituents), *De Nieuwe Taalgids* **74**, 197–205

Van den Toorn, M C 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is (Linking phonemes in compounds with loanwords as first constituents), *De Nieuwe Taalgids* **74**, 547–552

Van den Toorn, M C 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I)), *De Nieuwe Taalgids* **75**(1), 24–33

Van den Toorn, M C 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for the regulation of linking phonemes in nominal compounds II), *De Nieuwe Taalgids* **75**(2), 153–160

Van Jaarsveld, H and Rattink, G 1988, Frequency effects in the processing of lexicalized and novel nominal compounds, *Journal of Psycholinguistic Research* **17**, 447–473

Woordenlijst 1995, *Woordenlijst van de Nederlandse Taal 1995*, Sdu Uitgevers and Standaard Uitgeverij, 's-Gravenhage

Appendix

Materials for the Experiment: left constituent and right constituent (number of *s* responses, number of *en* responses, number of ∅ responses).

L1-R1: Left Constituent: abstract; Right Constituent: abstract:
taal staat (1, 7, 12); seizoen zang (11, 8, 1); loon gebrek (12, 2, 6); brand geluid (2, 3, 15); dienst toeval (2, 8, 10); vorm feest (2, 9, 9); symbool energie (4, 8, 8); naam toeslag (8, 4, 8); kracht maaltijd (8, 5, 7); contract opslag (9, 3, 8)

L1-R2: Left Constituent: abstract; Right Constituent: concrete-inanimate:
dienst vliegtuig (0, 8, 12); taal fles (1, 8, 11); seizoen jurk (17, 3, 0); symbool vork (3, 12, 5); loon altaar (3, 4, 13); vorm tapijt (3, 8, 9); brand nagel (4, 1, 15); kracht muts (5, 10, 5); naam standbeeld (5, 10, 5); contract telefoon (5, 4, 11)

L1-R3: Left Constituent: abstract; Right Constituent: concrete-human:
dienst consulent (0, 12, 8); taal heilige (0, 4, 15); brand leidster (1, 7, 11); seizoen zuster (15, 5, 0); contract producent (2, 4, 14); kracht idioot (2, 6, 12); symbool machinist (2, 6, 12); vorm redacteur (4, 4, 12); naam handelaar (5, 13, 2); loon violist (8, 1, 11)

L1-R4: Left Constituent: abstract; Right Constituent: concrete-animal:
dienst vogel (0, 3, 17); taal aap (0, 9, 11); symbool baars (1, 15, 4); loon uil (10, 0, 10); seizoen mees (15, 4, 1); vorm aal (3, 9, 8); contract gans (4, 4, 12); brand kat (4, 5, 11); naam slak (6, 8, 6); kracht os (8, 3, 9)

L2-R1: Left Constituent: concrete-inanimate; Right Constituent: abstract:
spier maaltijd (0, 10, 10); fiets staat (0, 11, 9); huis toeval (0, 11, 9); kaars energie (0, 14, 6); schoen geluid (0, 14, 6); arm gebrek (1, 14, 5); duim opslag (1, 7, 12); tand zang (2, 17, 1); trein feest (2, 6, 12); boot toeslag (3, 8, 9)

L2-R2: Left Constituent: concrete-inanimate; Right Constituent: concrete- inanimate:
kaars vork (0, 12, 8); tand fles (0, 15, 5); huis jurk (0, 3, 17); fiets vliegtuig (0, 8, 12); spier standbeeld (1, 10, 9); schoen nagel (1, 7, 12); arm tapijt (1, 9, 10); duim altaar (2, 11, 7); boot telefoon (2, 4, 14); trein muts (2, 8, 10)

L2-R3: Left Constituent: concrete-inanimate; Right Constituent. concrete- human: duim handelaar (0, 15, 5); tand producent (0, 15, 5); arm zuster (0, 18, 2); kaars idioot (0, 18, 2); trein redacteur (0, 3, 17); fiets machinist (0, 6, 14); huis consulent (0, 8, 12); spier violist (0, 8, 12); schoen heilige (1, 12, 7); boot leidster (2, 8, 10)

L2-R4: Left Constituent: concrete-inanimate; Right Constituent: concrete- animal: fiets vogel (0, 6, 14); huis baars (0, 6, 14); arm slak (0, 9, 11); kaars uil (1, 11, 8); spier os (1, 11, 8); trein gans (1, 7, 12); duim mees (2, 11, 7); schoen aal (2, 12, 6); tand aap (3, 12, 5); boot kat (8, 4, 8)

L3-R1: Left Constituent: concrete-animate; Right Constituent: abstract: weduwe toeslag (0, 13, 7); vis feest (0, 14, 6); marxist geluid (0, 19, 1); prins zang (0, 19, 1); koningin staat (0, 20, 0); christen energie (0, 8, 12); wees toeval (0, 9, 11); gast opslag (1, 14, 5); leerling maaltijd (1, 17, 2); vorst gebrek (2, 17, 1)

L3-R2: Left Constituent: concrete-animate; Right Constituent: concrete- inanimate: vis altaar (0, 13, 7); wees telefoon (0, 13, 7); gast vliegtuig (0, 16, 4); marxist tapijt (0, 19, 1); prins muts (0, 20, 0); christen jurk (0, 7, 13); weduwe standbeeld (1, 14, 5); vorst nagel (1, 18, 1); koningin fles (1, 19, 0); leerling vork (2, 16, 2)

L3-R3: Left Constituent: concrete-animate; Right Constituent: concrete-human: prins machinist (0, 14, 6); vis consulent (0, 6, 14); wees zuster (0, 6, 14); vorst heilige (1, 10, 9); gast idioot (1, 12, 7); leerling leidster (1, 16, 3); koningin producent (1, 17, 2); weduwe handelaar (1, 8, 11); marxist redacteur (2, 13, 5); christen violist (4, 6, 10)

L3-R4: Left Constituent: concrete-animate; Right Constituent: concrete-animal: gast baars (0, 13, 7); vorst slak (0, 13, 7); prins vogel (0, 18, 2); koningin gans (0, 20, 0); wees uil (0, 4, 16); vis aal (0, 5, 15); marxist mees (1, 16, 3); weduwe kat (1, 7, 11); christen os (3, 5, 12); leerling aap (6, 3, 11)

On-line selection

This chapter will be published as Andrea Krott, Robert Schreuder, and R. Harald Baayen: Linking elements in Dutch noun-noun compounds: constituent families as analogical predictors for response latencies, *Brain and Language*.

Abstract

This study addresses the choice of linking elements in novel Dutch noun-noun compounds. Previous off-line experiments (Krott, Baayen, & Schreuder, 2001, also chapter 2) revealed that this choice can be predicted analogically on the basis of the distribution of linking elements in the left and right constituent families, i.e. the set of existing compounds that share the left (or right) constituent with the target compound. The present study replicates the observed graded analogical effects under time-pressure, using an on-line decision task. Furthermore, the analogical support of the left constituent family predicts response latencies. We present an implemented interactive activation network model that accounts for the experimental data.

Introduction

Dutch noun-noun compounds often contain so-called linking elements or interfixes The two main ones are *-en-* and *-s-* as in *schaap+en+bout*, 'leg of mutton', or *schaap+s+kooi*, 'sheep fold' The linking *-en-* also occurs as the orthographic variant *-e-* Linguistic descriptions indicate that the occurrence of linking elements seems to be characterized by tendencies instead of clear-cut morphological rules (e g , Van den Toorn, 1982, Mattens, 1984, Haeseryn, Romijn, Geerts, Rooij, & Van den Toorn, 1997, see also Plank, 1976) A survey of the Celex Lexical Database (Baayen, Piepenbrock, & Gulikers, 1995) reveals that all phonological and morphological rules that are reported in the linguistic literature apply to only 51% of all Dutch compounds Of this subset, they correctly predict only 63%, which amounts to 32% of all compounds (Krott, Schreuder, & Baayen, in press, also chapter 3) Thus, rules do not provide an adequate account of linking elements Nevertheless, linking elements are used productively in novel compounds and, as it has been shown in Krott et al (2001, also chapter 2), with substantial agreement among native speakers

Whereas rule-based approaches have resulted in observationally inadequate analyses, an analogical approach has proved to be fruitful (Krott et al , 2001, also chapter 2, Krott et al , in press, also chapter 3) These studies, which used off-line production experiments in which participants had to choose the linking elements for novel Dutch compounds, report the crucial role of a graded, probabilistic factor the distribution of linking elements in what we have called the left and right constituent families The left (or right) constituent family is the set of existing compounds that share a left (or right) constituent with the novel compound We confirmed the predictive power of the constituent families by simulating the choice of linking elements by means of the analogical models AML (Skousen, 1989) and TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2000) In the case of the novel compounds used in our experiments, these models' choices were comparable to those of an average participant In the case of existing compounds, these models correctly predict 92% of the linking elements in all Dutch compounds in Celex, which is remarkable considering the mere 32% that can be accounted for by rules

In this paper we focus on three main questions First, do the left and right constituent families affect the choice of the linking element in Dutch novel noun-noun compounds when the choice has to be made under time-pressure? Second, do constituent families also affect the speed of the selection process? Third, can we formalize the processes that underly the choice and the response latencies in terms

of an implemented computational model?

In what follows, we first present an on-line production experiment in which responses have to be given under time-pressure The results show that the constituent families indeed also affect the choice of linking elements under time-pressure There is also an effect of the left constituent family on the reaction latencies We will give an interpretation of these findings in terms of a two-stage cognitive process

In the second part of the article, we present an interactive activation model that implements the morphological analogical processes A simulation study of the experimental results shows that our model can account for the effect of the constituent families on the choices as well as the response latencies

On-line production experiment

In order to come to grips with the influence of the constituent families on the choices of linking elements under time-pressure, we focus on the linking *-en-*

Method

Materials The materials were identical to those used in experiment 1 reported in Krott et al (2001, also chapter 2), i e three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3) The constituents of L1 and R1 had constituent families with as strong a bias as possible towards the linking element *-en-* Conversely, L3 and R3 showed a bias as strong as possible against *-en-* The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against *-en-*

As in the previous experiment, each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors Bias in the Left Position (Positive, Neutral, and Negative) and Bias in the Right Position (Positive, Neutral, and Negative) The items were presented to each participant in a separate random order

Procedure The participants performed an online cloze task The experimental items were presented on a computer screen as pairs of two compound constituents separated by two underscores We asked the participants to combine these constituents into new compounds and press as quickly as possible and according to the chosen linking element either a button labeled 'E/EN' or a button labeled 'S/-'

Participants were asked to give a sign when the pressed button was not intended We kept a protocol of these errors All participants pressed the EN-button with their dominant hand Each stimulus was preceded by a fixation mark in the middle of the screen presented for 500 ms After another 50 ms, the stimulus appeared in the same position and remained on the screen for 2000 ms The maximum time span allowed for the response was 2500 ms from stimulus onset Stimuli were presented on Nec Multicolor monitors in white lowercase 21 point Helvetica letters on a dark background The experiment lasted approximately 15 minutes

Occasionally, the first constituent may change its form when it is combined with a linking element (e g , *schip* ('ship') appears as *scheep* in the compound *scheepswerf* ('shipyard')) The instructions made clear that these changes were not of interest and could be ignored

Participants Twenty participants, undergraduates at Nijmegen University, were paid to participate in the experiment All were native speakers of Dutch

Results and discussion

We distinguished two different types of errors, time-out errors and self-corrections Taking both types of errors together, all participants performed the experiment with an error rate of maximal 10% and no item showed an error rate of more than 20% Therefore, all participants and items were included into further analyses Table 5 1 summarizes the percentages of *en* responses versus *not -en-* responses, the time-out errors and the self-corrections for the nine experimental conditions A by-item logit analysis (see, e g , Rietveld & Van Hout, 1993) of the valid responses showed a main effect of both Bias in the Left Position ($F(2,180) = 156\ 6$, $p < 001$) and Bias in the Right Position ($F(2,180) = 8\ 2$, $p < 001$) and no interaction between these factors ($F(4,180) = 4$, $p = 829$) Thus, the linking elements chosen by the participants follow both the Right and the Left Bias This is illustrated in the two upper left panels of Figure 5 1 for both the *-en-* and the *not -en-* responses With this result we have replicated the findings obtained with the off-line cloze task used in Krott et al (2001, also chapter 2) We conclude that the choice of the linking element for a novel compound is based on analogy even under time-pressure Apparently, the members of the constituent families become available quite fast

Note that participants responded slightly more often with *-en-* than expected on the basis of the bias Overall, more than half of the choices were *-en-* responses (1981 out of 3671, or 54%), leaving 46% for the other two linking elements Even though the experiment was designed to elicit an equal number of responses for both

Table 5.1: Mean percentages of selected linking elements and errors with varying Left and Right Bias for *-en-*. Left and Right Bias split up into the experimental conditions (Positive, Neutral, and Negative). en: *-en-* responses; not en: *not -en-* responses; self-corr: self-corrections; time-out: time-out errors. Standard deviations by items between parentheses.

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	en	90.0	(2.2)	92.4	(1.6)	82.1	(2.8)
	not en	8.8	(2.1)	6.7	(1.5)	16.4	(2.6)
	self-corr	1.0		1.4		1.4	
	time-out	1.2		1.0		1.4	
Neutral							
	en	68.1	(4.3)	75.5	(3.0)	57.9	(4.8)
	not en	30.0	(4.1)	22.1	(2.9)	39.8	(5.1)
	self-corr	1.2		0.7		1.4	
	time-out	1.9		2.4		2.4	
Negative							
	en	17.1	(3.5)	18.1	(3.3)	12.4	(2.9)
	not en	81.7	(3.5)	79.5	(3.5)	86.4	(3.1)
	self-corr	1.9		1.9		3.1	
	time-out	1.2		2.4		1.2	

push buttons, the push button for the 'not *-en-*' responses represents two linking elements instead of one. In the course of the experiment, participants may have become sensitive to *-en-* as being the most likely response. A similar response bias for *-en-* was present in the off-line cloze task reported by Krott et al. (2001, also chapter 2). An additional factor in the present on-line experiment may be that participants pressed the *-en-* push button always with their dominant hand.

A by-item logit analysis of the time-out errors revealed no effect, not of the Bias in the Left Position ($F(2,4) = 3.7$, $p = .124$) nor of the Bias in the Right Position ($F(2,4) = .8$, $p = .515$).

A by-item logit analysis of the self-corrections, on the other hand, revealed a reliable effect of the Bias in the Left Position ($F(2,4) = 11.2$, $p = .023$), but no effect of the Bias in the Right Position($F(2,4) = 3.7$, $p = .123$). Participants correct their choices more often if the left constituent has a bias against *-en-* than if it has a bias for *-en-*. This result becomes even more interesting when we take the direction of the self-correction into account, i.e. corrections from *-en-* to *not -en-*

or vice versa. Self-corrections occur almost exclusively when a participant has responded against the bias. A by-item logit analysis of the self-corrections from *-en-* to *not -en-* revealed a reliable effect of the Bias in the Left Position ($F(2,4) = 14.2$, $p = .015$), but again no effect of the Bias in the Right Position ($F(2,4) = 3.2$, $p = .150$). A stepwise by-item logit analysis of the self-corrections from *not -en-* to *-en-* also revealed a reliable effect of the Bias in the Left Position only ($F(2,6) = 5.8$, $p = .039$).

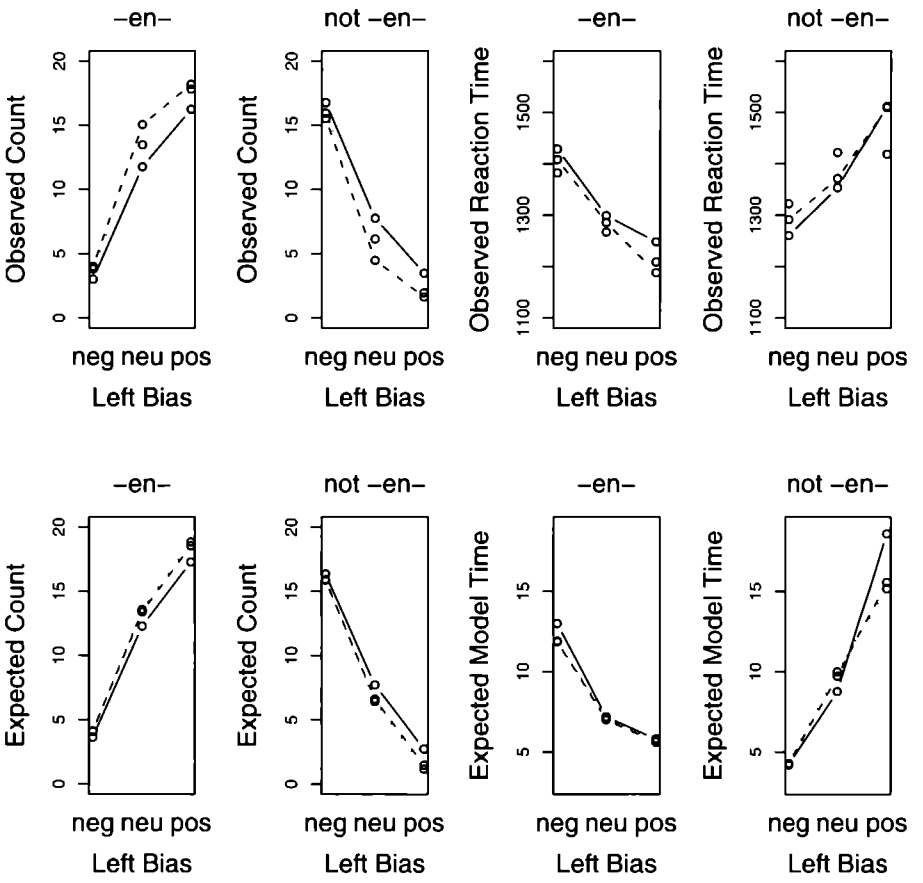


Figure 5.1: Interaction plots for the observed and expected counts and response latencies of the linking element *-en-* on the one hand, and the other two linking elements *-s-* and *-0-* (= *not -en-*) on the other hand, with the left constituent bias on the horizontal axis, and the right constituent bias indicated by line type (solid line: negative bias; dashed line: neutral bias; dotted line: positive bias).

Table 5 2 Mean Response latencies for varying Left and Right Bias for *-en-* Left and Right Bias split up into the experimental conditions (Positive, Neutral, and Negative) Standard deviations by items between parentheses

Left Position		Right Position					
		Positive		Neutral		Negative	
Positive							
	RT en	1209	(130)	1188	(122)	1248	(129)
	RT not en	1419	(611)	1509	(799)	1512	(310)
Neutral							
	RT en	1267	(124)	1286	(145)	1299	(149)
	RT not en	1422	(458)	1371	(179)	1354	(152)
Negative							
	RT en	1382	(592)	1408	(491)	1429	(672)
	RT not en	1322	(176)	1291	(177)	1260	(159)

Table 5 2 shows the mean response latencies (calculated for the valid responses) for the nine experimental conditions An analysis of variance of the *-en-* and *not -en-* responses revealed a main effect of the Bias in the Left Position (*-en-* responses $F_1(2,180) = 15\,2$, $p < 001$, $F_2(2,180) = 16\,3$, $p < 001$, *not -en-* responses $F_1(2,180) = 10\,7$, $p < 001$, $F_2(2,180) = 10\,8$, $p < 001$), but no effect of the Bias in the Right Position (*-en-* responses $F_1(2,180) = 7$, $p = 519$, $F_2(2,180) = 8$, $p = 462$, *not -en-* responses $F_1(2,180) = 1\,5$, $p = 237$, $F_2(2,180) = 9$, $p = 915$) Apparently, the Right Bias does have influence on the choice of the linking element, but not on the response latency The upper two right panels of Figure 5 1 show the effect of the Left Bias on the reaction latencies for both *-en-* and *not -en-* responses Participants react faster when the response follows the bias than when the response conflicts with the bias

We also tested whether the Left and Right Bias of the preceding experimental trial and the choice made for that trial had an influence on the choice, in addition to the effects of the Left and Right Bias A logit analysis that included the preceding Left and Right Bias and the preceding choice along with the Left and Right Bias themselves revealed a significant effect only for the Left and Right Bias, both with respect to the choices and with respect to the response latencies

Summing up, we replicated the finding that linking elements in novel Dutch compounds are chosen on the basis of analogy As in Krott et al (2001, also chapter 2),

both the bias of the left constituent family and the right constituent family show a main effect on the choice. The left constituent family also plays a crucial role for the response latencies: Responses that follow the bias require less processing time. The right constituent family, however, that already revealed a weaker effect on the choices, does not predict the response latencies.

What kind of cognitive processes might account for these findings? In order to explain the absence of an effect of the right constituent family on the reaction times, we propose to distinguish between an early selection process and a series of processing stages during which activation accumulates up to response initiation. In the early selection process, a linking element is chosen based on maximum likelihood, i.e. on the distribution of linking elements in both the left and right constituent families. Along the lines of the interactive activation model that has been outlined in Krott et al. (2001, also chapter 2), we hypothesize that the lemma representations of the constituents of the novel compound activate the corresponding left and right constituent families. The compounds in these families then activate the linking elements they contain. Since the left constituent family has a stronger effect than the right constituent family, we assume that the members of the left constituent family are initially higher activated than the members of the right constituent family. The higher activation of the left constituent family implies that the linking elements receive more activation from members of the left constituent family. After the initial activation of linking elements, the activation flows back and forth between the linking elements and the constituent families. The activation accumulates until the selected linking element has become sufficiently activated to reach an awareness threshold, which initiates the response. We hypothesize that the alternating activation flow between the constituent families and the linking elements leads to an exponential increase of the activation of the already higher activated members of the left constituent family and a comparably slow increase of activation of the member of the right constituent family. This results in response latencies that appear to be based solely on the bias in the left constituent family, the relatively weak contribution of the right constituent family being masked.

An interactive activation model

Introduction

In previous studies, we used AML (Skousen, 1989) and TiMBL (Daelemans et al., 2000) as analogical tools to model the choice of linking elements in novel compounds (Krott et al., 2001, also chapter 2; Krott et al., in press, also chapter 3). The selection of a linking element can be understood as a classification problem, and both these models are very much suited to this task. However, they are restricted in that they are not designed to model response latencies. We therefore decided to develop a symbolic activation model that incorporates, in part, aspects of TiMBL.

Figure 5.2 illustrates the connectivity structure for a simple lexicon with ten compounds for the situation in which the novel compound *schaap-?-oog* (sheep's eye) has been conceptualized, with *schaap* in the modifier position (LEFT) and *oog* in the head position (RIGHT). As outlined in the previous section, initially, activation flows from the lemma representation of *schaap* to the wordforms (the lexemes in the sense of Levelt, 1989) with which it is connected, modified by the (identical) weights w_1 (model parameter: IG-weight left constituent, γ_1). Similarly, activation flows from the lemma representation of *oog* to the wordforms of the compounds in which *oog* is the head, modified by the (identical) weights w_2 (model parameter: IG-weight right constituent, γ_1). The weight w_1 is larger than the weight w_2, in accordance with the empirical finding that the left constituent family has greater analogical force than the right constituent family. Only members of the two constituent families are activated. Therefore, compounds such as the members of the left constituent family of *lam* (see Figure 5.2) are not activated. From the activated wordforms, activation flows further to the linking elements. The wordforms with the linking element *-en-* support the linking element *-en-*, similarly, the wordforms with the linking elements *-s-* and *-∅-* support the linking elements *-s-* and *-∅-*, respectively. The linking element that receives the highest activation from the wordforms is the linking element that is most likely to be selected. Following selection, activation flows back from the linking elements to the wordforms, and from the wordforms to the lemma representations. The forward activation flow from the lemmas to the linking elements, and the backward activation flow from the linking elements to the lemmas, jointly constitute one resonance cycle. Generally, a series of resonance cycles, the time steps of the model, are required for a selected linking element to become sufficiently activated to reach the level of awareness required for response execution.

Apart from the weights for the left and right constituents, the model contains some other parameters: The general decay δ determines the activation decay of nodes in the network. The resonance weight ρ specifies the strength of the activation resonance, while the activation is only passed on from compounds whose activation exceeds a similarity threshold ϑ. The overall bias for *-en-* that has been observed in the experiment can be adjusted by changing the parameter β. The strength of the bias increases if the parameter ξ has a value above zero. Furthermore, one can specify whether the frequency of the compounds should affect the activation increase. A linking element reaches awareness once its activation reaches a threshold θ. In order to guarantee that the model terminates, the number of maximal time steps has to be set. In the following subsection, we explain the model's details. The reader may skip that part without loosing the main thread of the argument.

Technical details

The connectivity structure of the model is defined formally by means of two matrices, $\mathbf{C}$ and $\mathbf{E}$. Let $\mathbf{C}$ denote the connectivity matrix of n_u wordforms and n_f feature-value pairs:

$$\mathbf{C} = \begin{pmatrix} c_{1\,1} & c_{1\,2} & \cdots & c_{1\,n_f} \\ c_{2\,1} & c_{2\,2} & & c_{2\,n_f} \\ \vdots & \vdots & & \vdots \\ c_{n_u\,1} & c_{n_u\,2} & \cdots & c_{n_u\,n_f} \end{pmatrix}, \tag{5.1}$$

with

$$c_{u\,F} = \begin{cases} 1 & \text{if wordform } u \text{ is connected to feature } F, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

In the present working example (Figure 5.2), n_u = 10 and n_f = 2. The relevant features are the left and right constituent positions (modifier and head), the values of these features are the lemma representations of *schaap* and *oog* respectively. Similarly, let $\mathbf{E}$ denote the connectivity matrix of the n_u words with the n_e exponents (the three linking elements studied here):

$$\mathbf{E} = \begin{pmatrix} e_{1\,1} & e_{1\,2} & \cdots & e_{1\,n_e} \\ e_{2\,1} & e_{2\,2} & \cdots & e_{2\,n_e} \\ \vdots & \vdots & & \vdots \\ e_{n_u\,1} & e_{n_u\,2} & & e_{n_u\,n_e} \end{pmatrix}. \tag{5.3}$$

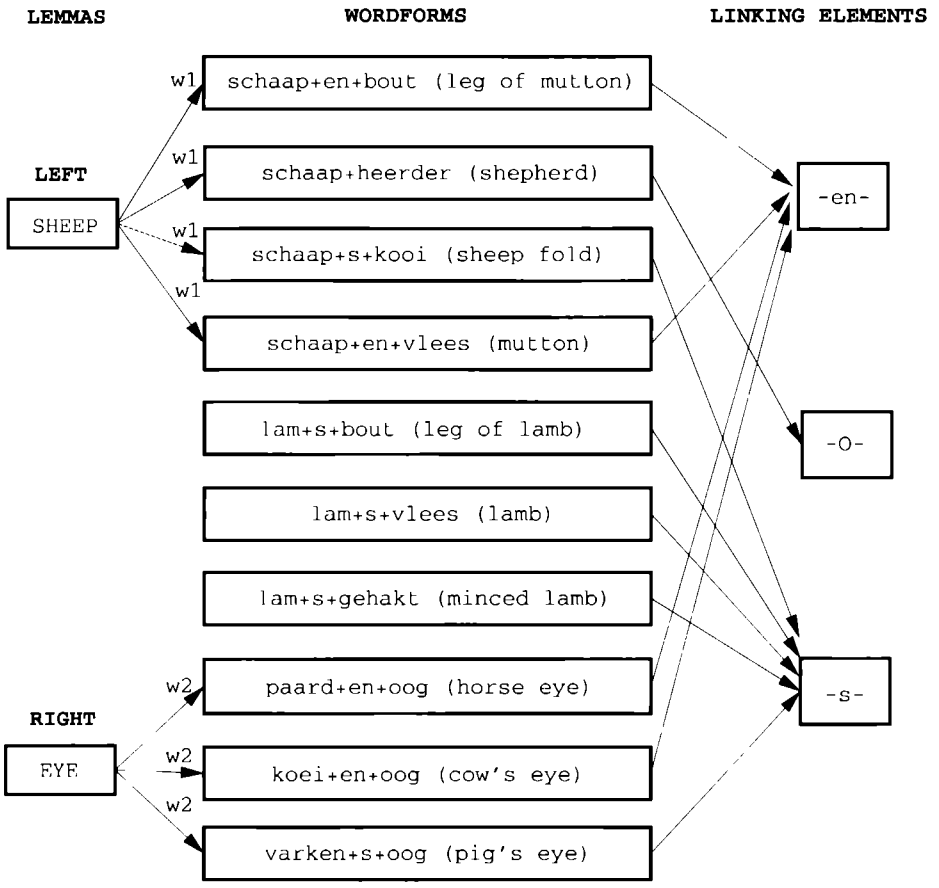


Figure 5.2: Connectivity of a simple lexicon: lemmas (left layer), wordform representations (lexemes in the sense of Levelt (1989), central layer), and linking elements (right layer).

with

$$i_{w,e} = \begin{cases} 1 & \text{if wordform } w \text{ is connected to exponent } e, \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

These two matrices completely define the connectivity in the model. For the present example, these two matrices have the form

$$\mathbf{C} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{E} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.5}$$

Note that the connectivity matrix $\mathbf{C}$ differs for each pair of left and right target constituents. For every such pair, we consider only that section of the lexical connectivity that is relevant for precisely this pair of constituents.

The forward activation flow from the lemmas to the linking elements is co-determined by the weights on the connections between the lemmas and the wordforms, as well as by the frequencies of these wordforms. Let γ denote the vector of feature information gain weights in the sense of Daelemans et al. (2000),

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n_F} \end{pmatrix}, \tag{5.6}$$

with

$$\gamma_i = w_i = H(\mathcal{E}) - H_i'(\mathcal{E}). \tag{5.7}$$

To understand this equation, let $F_i \in \mathcal{F}$ denote the i-th feature, and let this feature assume values $F_{ij}, j = 1, 2, \ldots, c(F_i)$, with $c(F_i)$ the cardinality of the set of values that F_i can assume. In the present working example, $c(F_1)$ = 5 as there are 5 different left constituents in the lexicon, and $c(f_2)$ = 6. Furthermore, let $e_i \in \mathcal{E}$, $i = 1, 2, \ldots, c(\mathcal{E})$, with $c(\mathcal{E})$ the cardinality of $\mathcal{E}$, denote the i-th exponent. In our work-

ing example, we have three exponents, hence $c(\mathcal{E}) = 3$. The entropy of E equals

$$H(\mathcal{E}) = -\sum_{i=1}^{c(\mathcal{E})} P(e_i) \log_2 P(e_i), \tag{5.8}$$

with $P(e_i)$ the relative frequency of the i-th linking element among the wordforms. The entropy of $\mathcal{E}$ is reduced by introducing knowledge of the value of feature F_i. The weighted entropy of $\mathcal{E}$ given knowledge of the value of F_i is

$$H_i'(\mathcal{E}) = \sum_{j=1}^{c(F_i)} P(F_{ij}) H(\mathcal{E}|F_{ij}), \tag{5.9}$$

with $P(F_{ij})$ the relative frequency of the j-th value of F_{ij} among all the values that feature F_i assumes, and with $H(\mathcal{E}|F_{ij})$ the entropy calculated over those exponents that are linked with wordforms sharing the j-th value of feature F_i. Thus, the information gain weight of feature i can be understood as the reduction in entropy achieved by introducing knowledge of the value of feature F_i. Note that all connections from the modifier position share the same weight, the information gain weight of the left constituent, and that likewise the connections from the head position share the same information gain weight. All information gain weights are easily estimated on the basis of the wordforms in the lexicon. No training of the model is required.

In addition to the connection weights, the model makes use of a vector φ of word frequency weights:

$$\boldsymbol{\varphi} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{n_u} \end{pmatrix}. \tag{5.10}$$

The frequency weight φ_i is a function ϕ of the Celex frequency f_i of wordform w_i:

$$\phi(f_i) = \frac{1\,0}{1\,0 + \log(f_i)}. \tag{5.11}$$

Inverse frequency weighting favors the analogical contribution of the lower-frequency words, the words that most clearly express the regularities in the lexicon (cf. Baayen & Sproat, 1996). It is in symmetric contrast with the non-inverse frequency effect that arises when wordforms directly feed articulation (Jescheniak & Levelt, 1994).

The pattern of activation values of the wordforms after the first forward pass of

activation,

$$\mathbf{s} = (\mathbf{C} \cdot \boldsymbol{\gamma}) * \boldsymbol{\varphi} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n_w} \end{pmatrix}, \tag{5.12}$$

is a vector of by-wordform similarity scores. Each similarity score specifies how much activation a given wordform will pass on to the exponent with which it is connected. By applying a thresholding function Θ, we obtain the equivalent of the standard k-NN distance sets, but now defined in terms of similarities instead of distances:

$$\Theta(s_i, \vartheta) = \begin{cases} s_i & \text{if } s_i >= \vartheta \\ 0 & \text{otherwise,} \end{cases} \tag{5.13}$$

with ϑ representing a similarity threshold. In the present simulation, the value of ϑ is set to zero. In other words, we have allowed even distant neighbors to co-determine the selection of the linking elements. But by choosing an appropriate value for ϑ, only those words that are sufficiently similar to the target input affect the activation of the linking elements.

The activation of the wordforms is passed on to the exponents. The vector of activations of the exponents $\mathbf{e}$ after the first forward pass of activation has run its course equals

$$\mathbf{e} = \mathbf{E}^T \cdot \mathbf{s} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n_e} \end{pmatrix}. \tag{5.14}$$

The probability of selecting the i-th linking element is

$$P(i) = \frac{e_i}{\sum_{j=1}^{n_e} e_j}. \tag{5.15}$$

When no frequency weighting is used, the resulting probabilities of the linking elements are identical to those obtained by applying the k-NN nearest neighbor algorithm with information gain weighting as developed in TiMBL.

Maximum likelihood selection according to (5.15) allows us to model the selection of the linking elements, but not the time required for executing an actual response. As the constituent family of the right constituent affected the choice of the linking elements but not the response latencies, we need a mechanism that introduces noise in such a way that the strongest factor, the left constituent family, masks the

effect of the weaker factor, the right constituent family In the present model, this is accomplished by means of resonance in the network We assume that this resonance leaves activation traces, either in the connections, or in the activation levels of the wordforms As the wordforms themselves are not the forms to be produced, we prefer to view the activation traces as accumulating in the connections However, the following formal definition is neutral with respect to these interpretations

We assume that the activation received by the wordforms from the lemmas during the initial forward pass of activation leaves an activation trace in the network of connections between the lemma layer and the wordforms, proportional to what we call the forward activation matrix $\mathbf{F}$

$$\mathbf{F} = (1 + \mathbf{s}) * \mathbf{C} \tag{5 16}$$

Following maximum likelihood selection of a linking element, activation flows back from the exponents to the lemma layer, again increasing the activation in this network of connections, this time proportional to what we call the backward activation matrix $\mathbf{B}$, indexed here for the initial time step $t = 1$

$$\mathbf{B}_1 = (\mathbf{E} \quad \mathbf{e}) * \mathbf{C} \tag{5 17}$$

Let $\mathbf{A}_t$ denote the activation pattern at time step t, $t = 0\ 1\ 2$, with $\mathbf{A}_0 = \mathbf{C}$ For $t = 1$, the first resonance cycle, we define

$$\mathbf{A}_1 = \delta(\mathbf{A}_0 + \rho(\mathbf{F} + \mathbf{B}_1)) \tag{5 18}$$

with δ a general activation decay, and ρ a resonance weight, a parameter allowing us to specify the granularity of the resonance The state of the model at an arbitrary time step t is, in summary form

$$\begin{aligned}
\mathbf{s}_t &= (\mathbf{A}_{t\ 1}\ \gamma) * \varphi \\
\mathbf{e}'_t &= \mathbf{e}_{t-1} + \mathbf{E}^{l}\ \mathbf{s}_t \\
\mathbf{B}_t &= (\mathbf{E}\ \mathbf{e}'_t) * \mathbf{C} \\
\mathbf{A}_t &= \delta(\mathbf{A}_{t\ 1} + \rho(\mathbf{F} + \mathbf{B}_t)) \\
&= \delta(\mathbf{A}_{t-1} + \rho(([1 + \mathbf{s}_1] + [\mathbf{E}\ \mathbf{e}'_t]) * C)) \\
\mathbf{e}_t &- \mathbf{e}'_t + \mathbf{b}
\end{aligned} \tag{5 19}$$

The last line specifies that the activation of the linking elements is modified by the

vector $\mathbf{b}$. This vector allows us to implement the observed response bias for the *-en-* linking element,

$$\mathbf{e}_t = \begin{pmatrix} e_{\text{-en-}} \\ e_{\text{-}\emptyset\text{-}} \\ e_{\text{-s-}} \end{pmatrix} + \begin{pmatrix} \beta\xi^{t-1} \\ 0 \\ 0 \end{pmatrix}, \tag{5.20}$$

with $\beta > 0$ and $\xi >= 1$. Note that this bias for *-en-* increases during the resonance cycles when $\xi > 1$. In other words, we assume that the response bias is a task factor that is itself external to the connectivity in the lexical network.

A selected linking element reaches awareness once its activation has reached a present threshold value θ. The time step at which this threshold is reached is taken to represent the model's response latency. Model times exceeding a preset time limit are not taken into account, just as response latencies exceeding the time-out limit are not taken into account.

Simulation results

A reasonable fit of this model to the present experimental data was obtained with the following parameter values: IG-weight left constituent: γ_1 = 1.12; IG-weight right constituent: γ_2 = .10; general decay δ = .97; resonance weight ρ = .05; activation threshold θ = 100.0; *-en-* bias parameters β = 2.5 and ξ = 1.2, with timeout after 25 time steps, with frequency weighting and no similarity threshold (ϑ = 0). Figure 5.1 presents a visual summary of the goodness of fit, and Table 5.3 shows that the same main effects that can be observed for the experimental data also emerge in the simulation. The same holds for the interaction term for left and right constituent bias, except for the logit analysis of the observed and expected counts. The model suggests a minor interaction that does not receive clear support from the empirical data. However, given that the model has no sources of variation other than those provided by the constituent families, this small interaction, that qualitatively is of the same kind as the non-reliable interaction visible in the empirical results, is not a source of serious concern. We conclude that our morphological resonance model provides a reasonable first approximation of the role of analogical cognition in the production of Dutch noun-noun compounds.

Table 5.3: Goodness of fit statistics: a logit analysis of the observed and expected counts, and analyses of variance for the reaction times corresponding to the *-en-* and the *not -en-* responses.

Logit Analysis of Counts

	Observed		Expected	
Left Bias	$F(2,180) = 156.6$	$p < .001$	$F(2,180) = 902.99$	$p < .001$
Right Bias	$F(2,180) = 8.2$	$p < .001$	$F(2,180) = 12.11$	$p < .001$
Interaction	$F(4,180) = .37$	$p = .829$	$F(2,180) = 2.76$	$p = .029$

Analysis of Variance of log RT: *-en-*

	Observed		Expected	
Left Bias	$F(2,169) = 16.3$	$p < .001$	$F(2,180) = 177.89$	$p < .001$
Right Bias	$F(2,169) = .8$	$p = .462$	$F(2,180) = .68$	$p = .510$
Interaction	$F(4,169) = .1$	$p = .969$	$F(2,180) = .19$	$p = .943$

Analysis of Variance of log RT: *not -en-*

	Observed		Expected	
Left Bias	$F(2,166) = 10.8$	$p < .001$	$F(2,147) = 165.14$	$p < .001$
Right Bias	$F(2,166) = .9$	$p = .915$	$F(2,147) = .04$	$p = .978$
Interaction	$F(4,166) = .4$	$p = .437$	$F(2,147) = .90$	$p = .468$

General discussion

In this study we addressed three related questions. First, does the distribution of linking elements in the right and left constituent families predict the choice of the linking elements in novel compounds not only in an off-line cloze task but also in a speeded decision task? Second, does this distribution also predict the speed with which these decisions are made? Third, is it possible to model the processes involved in the on-line experiment in a psycholinguistically plausible way?

The on-line experiment that we presented in this study showed that indeed the effect of the left and right constituent families on the choice of linking elements in Dutch noun-noun compounds also occurs under time-pressure. This effect is not restricted to the choices made by the participants, it also emerges in their response latencies. We observed an asymmetry between the choice pattern and the reaction time pattern, however. Both the left and the right constituent families play a role for the choices, while for the response latencies it is only the left constituent family that is a predictor.

We interpreted these results in terms of a two-stage cognitive process. In the first

stage, a linking element is selected on the basis of a maximum likelihood selection following initial activation spreading from the left and right constituent families to the linking elements In the second stage, the activation of the selected linking element increases until it reaches an awareness threshold, after which the selected response can be initiated We assume that in this process the relatively weak effect of the right constituent is masked by the additional variability of this second processing stage

We have made this explanation more explicit by means of a computational simulation model In this model, the first processing stage is captured by a spreading activation mechanism that is mathematically equivalent to a k-NN nearest neighbor classifier as used in machine learning approaches to natural language processing (e g , Daelemans et al , 2000) The second processing stage is captured by allowing activation to resonate in the lexical network

A simulation study of the results of our experiment showed that our model can account for the analogical effects on both the choices and the response latencies An advantage of the present psycholinguistic model compared to linguistic models of analogy such as AML and TiMBL is that it captures, within a spreading activation framework, the patterns in the data not only with respect to the choices but also with respect to the reaction times

The results that we have obtained are difficult to account for within a traditional approach based on symbolic rules As mentioned in the introduction, the rules that have been formulated for the linking elements in Dutch have insufficient predictive power (Krott et al , in press, also chapter 3) Given the syntagmatic nature of rules, this lack of predictive power is not so surprising By definition, symbolic rules do not have access to constituent families They may be sensitive to particular properties of left and right constituents, for instance, to whether the first constituent ends in a vowel In order to capture generalizations, rules can only be sensitive to properties of words, and not to specific words

Interestingly, the phenomenon that we have studied here is not syntagmatic in nature, but paradigmatic The left and right constituent families both constitute positional paradigms In fact, each such paradigm constitutes its own domain of markedness A positive bias for *-en-* as linking element indicates that this linking element is the locally unmarked form

The notion of local markedness as introduced by Tiersma (1982) concerns the fact that some marked forms may behave as unmarked forms For instance, noun plurals denoting objects that naturally occur in pairs or groups (e g , 'eyes', 'sheep')

may serve as attractors in language change, a role that is normally reserved for the unmarked singular forms of words such as 'nose' and 'nightingale'. Not surprisingly, locally unmarked plurals are much more frequent than their corresponding singulars than marked plurals, which tend to be less frequent than their singulars. They are also conceptually more central than their singulars. Although linking elements lack this conceptual aspect, they share the property of being locally unmarked with plural forms such as 'eyes'. Just as 'eyes' occurs, for the domain of the lemma EYE, more often than the singular 'eye', a locally unmarked linking element with a large positive bias in the relevant constituent family occurs more often than the other linking possibilities. For the local domains of constituent families, the formally unmarked linking element -∅-, which also occurs in the majority (69%) of Dutch compounds, may be rare and if so, locally marked. Furthermore, markedness and the constituent family bias have in common that they are both graded in nature.

Finally, markedness theory claims that unmarked forms are easier to process than marked forms (Dressler, Mayerthaler, Panagl, & Wurzel, 1987). Given that the left constituent families constitute independent markedness domains, the shorter response latencies of the locally unmarked linking elements, the dominant linking elements in their own local markedness domains, is exactly as expected. From a methodological point of view, it is interesting to find that classic structuralist notions such as markedness and paradigmatics can help to understand a graded analogical phenomenon such as the realization of linking elements in Dutch noun-noun compounds.

References

Baayen, R. H. and Sproat, R.: 1996, Estimating lexical priors for low-frequency morphologically ambiguous forms, *Computational Linguistics* **22**, 155–166.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 2000, TiMBL: Tilburg Memory Based Learner Reference Guide. Version 3.0, *Technical Report ILK 00-01*, Computational Linguistics Tilburg University.

Dressler, W., Mayerthaler, W., Panagl, O. and Wurzel, W.: 1987, *Leitmotifs in Natural Morphology*, Benjamins, Amsterdam.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Jescheniak, J. D. and Levelt, W. J. M.: 1994, Word frequency effects in speech production: Retrieval of syntactic information and of phonological form, *Journal of Experimental Psychology: Learning, Memory and Cognition* **20**(4), 824–843.

Krott, A., Baayen, R. H. and Schreuder, R.: 2001, Analogy in morphology: modeling the choice of linking morphemes in Dutch, *Linguistics* **39**(1), 51–93.

Krott, A., Schreuder, R. and Baayen, R. H.: in press, Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds, *in* R. Skousen (ed.), *Analogical Modeling: An Exemplar-Based Approach to Language*.

Levelt, W.: 1989, *Speaking. From intention to articulation*, The MIT Press, Cambridge, Mass.

Mattens, W. H. M.: 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343.

Plank, F.: 1976, Morphological aspects of nominal compounding in German and certain other languages: what to acquire in language acquisition in case the rules fail?, *in* G. Drachman (ed.), *Akten des 1. Salzburger Kolloquiums über Kindersprache*, number 2 in *Salzburger Beiträge zur Linguistik*, Gunter Narr, Tübingen, pp. 201–219.

Rietveld, T. and Van Hout, R.: 1993, *Statistical Techniques for the Study of Language and Language Behaviour*, Mouton de Gruyter, Berlin.

Skousen, R.: 1989, *Analogical Modeling of Language*, Kluwer, Dordrecht.

Tiersma, P. M.: 1982, Local and General Markedness, *Language* **58**, 832–849.

Van den Toorn, M. C.: 1982, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I)), *De Nieuwe Taalgids* **75**(1), 24–33.

German linking elements

This chapter has been submitted as Andrea Krott, Robert Schreuder, R. Harald Baayen and Wolfgang U. Dressler: Analogical effects on linking elements in German compounds.

Abstract

This paper focuses on the factors determining the selection of linking elements in German noun-noun compounds. A previous study by Dressler, Libben, Stark, Pons & Jarema (2001) presents evidence mainly for the effect of the category of the left constituent, but also for analogical effects of existing compounds sharing the left constituent with the target compound, the left constituent family. In the case of Dutch linking elements, Krott, Baayen & Schreuder (2001, also chapter 2) report evidence for paradigmatic effects of both the left and the right constituent families. The present study investigates in more detail the possible paradigmatic analogical effect of the left and right constituent families on linking elements in German compounds. We present three production experiments that confirm the effect of the left, but not of the right constituent family on the three main German linking possibilities: *-s-*, *-(e)n-*, and *-∅-*. Simulation studies of the responses in our experiments, using a computational model of paradigmatic analogy, reveal that both the left constituent and its phonological and morphological properties, notably its rime, its gender, and its inflectional class, simultaneously codetermine the selection of a linking element. We interpret the results as the combined effect of different kinds of analogical similarities and we outline a symbolic interactive activation model that merges these analogical effects in one psycholinguistically motivated processing mechanism.

Introduction

A phenomenon occurring in various languages across different language families is the insertion of linking elements between the immediate constituents of noun-noun compounds. With respect to predictability, such linking elements vary. In English, a linking *-s-* can be found in frozen forms like *hunt+s+man*, *state+s+man*, *lamb+s+wool*, or *grand+s+manship*. These forms are exceptional and have to be stored item by item in the lexicon. In other languages, however, linking elements are either fully predictable or partly predictable. A language with fully predictable linking elements is, for instance, Russian. Russian root-root compounds contain *-o-* when the first root ends in a hard consonant as in *par-o-voz* (steam-O-carry 'locomotive'), after a soft consonant they contain *-e-* as in *pyl-e-sos* (dust-E-suck 'vacuum cleaner') (Unbegaun, 1967). Such fully predictable linking elements are easily accounted for in terms of general syntagmatic rules.

Linking phenomena in compounds that are partly predictable can be found in Germanic languages such as German, Danish, Dutch, Afrikaans, Swedish, and Norwegian. Previous research (Krott et al., 2001, also chapter 2) has shown that the selection of the linking elements *-s-* and *-en-* that occur in Dutch noun-noun compounds (e.g., *schaap+s+kooi* 'sheep fold' and *boek+en+kast* 'book shelf') can be explained on the basis of paradigmatic analogy. The strongest predictor of Dutch linking elements is the distribution of linking elements in the set of compounds that share the left constituent with the target compound, a paradigmatic set that we refer to as the left constituent family. For instance, the choice of the linking element for the novel compound *schaap-?-oog* (sheep eye) is based on the linking elements in compounds such as (1).

(1) *schaap+en+bout* 'sheep leg'
schaap+en+tong 'sheep tong'
schaap+en+wol 'lambs wool'
schaap+s+kooi 'sheep fold'
schaap+herder 'shepherd'

There is also evidence for a somewhat smaller paradigmatic effect of the right constituent family, i.e. the set of compounds that share the right constituent with the target compound. Thus, the realization of the linking element in *schaap-?-oog* is co-determined by compounds such as (2), a right constituent family without clear bias for a particular linking element. Because of the strong effect of the left con-

stituent family, *schaap-?-oog* would most probably become *schaap-en-oog*.

(2) *varken+s+oog* 'pig eye'
kip+en+oog 'chicken eye'
kunst+oog 'artificial eye'

Other studies have focused on the effect of the preceding suffix and of the preceding rime on the linking element in Dutch compounds (Krott et al., 2001, also chapter 2, Krott, Schreuder, & Baayen, in press-a, also chapter 3). Although these factors also play a role for Dutch, they are typically overruled by the paradigmatic effect of the left constituent family. In addition to these form effects on Dutch linking selection, a semantic effect of the class of the left constituent has been observed. For instance, left abstract constituents are often combined with the linking *-s-* (Krott, Krebbers, Schreuder, & Baayen, in press, also chapter 4). In contrast to this paradigmatic analogical approach, that explains both the occurrence of linking elements in existing compounds and their selection in novel compounds, a strict syntagmatic rule-based analysis appears to be observational inadequate. The rules for Dutch linking elements that have been proposed in the literature (e.g., Van den Toorn, 1982a, 1982b; Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997) do not capture all possible contexts in which linking elements can occur. Moreover, taking the subset of compounds of the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) to which the rules are applicable, their prediction accuracy is a disappointing 63%, i.e. 32% of all compounds in CELEX (Krott et al., 2001, also chapter 2; Krott et al., in press-a, also chapter 3).

In the present paper, we turn to another language with partly-predictable linking elements, German. The main German non-Latinate linking elements are *-s-*, *-e-*, *-n-*, *-en-*, *-ens-*, *-es-*, and *-er-*. In addition, the linking elements *-e-* and *-er-* can trigger umlaut in a preceding umlautable vowel. Most of the noun-noun compounds, however, namely 65% of the noun-noun compounds in the CELEX lexical database, do not contain any linking elements. This is slightly less than the 69% Dutch noun-noun compounds that occur without any linking element. The most frequent German linking element is the linking *-s-*, which occurs in 17% of all compounds, followed by *-(e)n-* with 15%. The remaining linking elements occur rarely (*-es-*: 1.5%; *-e-*: 1%; *-er-*: 0.4%; *-ens-*: 0.2%). As in Dutch, German linking elements have their diachronic origin in earlier inflectional forms (see Dressler & Merlini Barbaresi, 1991; Fuhrhop, 1996). The system of German noun inflections is much more complex than the Dutch system, mirroring the substantial difference of the

respective inflectional systems.

Occasionally, the first constituent of a German compound may change its form when it is combined with a linking element (e.g., *Hand* 'hand' appears as *Händ* in the compound *Händ+e+Druck* 'handshake'). It is also possible that the left constituent is shortened when it appears in a compound (e.g., *Firma* 'company' in *Firm+en+Name* 'company name' or *Farbe* 'color' in *Farb+Fernseher* 'color television'). It has been proposed that *Hände* in *Händedruck* should not be analyzed as *Händ* followed by the linking element *-e-*, but as two constituents forming an independent unit that serves as a compounding stem form (Fuhrhop, 1998). We will come back to this issue in the general discussion.

A recent experimental study by Dressler, Libben, Stark, Pons & Jarema (2001) considered the question whether the choice of German linking elements is governed by rules or analogy. Dressler et al. introduce ten linguistic categories of left constituents based on grammatical gender, phonological form, and inflectional class. These categories differ in the choice of linking elements. For instance, one of these categories comprises schwa-final feminine nouns which always occur with a linking *-n-* as in *Suppe+n+Topf* 'soup+LINK+pot'. The authors proceed by determining the appropriate linking elements on the basis of eight (once six) exemplars for each of these ten categories. For the actual experiment, they selected three left constituents of each category for presentation. The task of the experiment was to create novel compounds. Although most of the responses were well predicted by the categories and their appropriate linking elements, the category of root-based concatenation with truncation of the word-final Sg. schwa of a feminine (e.g., *Sprache* in *Sprach-labor* 'language laboratory') revealed an unexpected number of responses that deviate from the expected linking element. Dressler et al. assume that this variation is due to an analogical effect of the existing compounds that share the first constituent with the target compound, i.e. the left constituent family. Interestingly, this category is not the only one that revealed variation. For instance, the left constituent *Stern* led to 57% *-en-* responses, which is the expected linking element, but also to 43% *-∅-* responses. Interestingly, 27% of the members in the constituent family of *Stern* in the CELEX contain a linking *-en-*, while 73% contain a *-∅-*. Even if these percentages would lead the distribution into another direction, the fact that both linking elements occur as responses again hints at an analogical effect of the left constituent family.

The aim of the present study then is to investigate in more detail the possible paradigmatic analogical effect of the constituent families on the selection of linking

elements in novel German compounds. Note that the idea that analogy might be involved in the formation of German compounds has already been suggested by Becker (1992). However, he makes use of a general fuzzy notion of analogy that contrasts with the computationally tractable paradigmatic analogy with which we are concerned here.

In what follows, we present three production experiments that test the effect both of the left and the right constituent families on the three main German linking possibilities: *-s-*, *-(e)n-*, and *-∅-*. These linking elements occur often enough in compounds to provide a substantial set of experimental items. For all three production experiments that are presented in this paper, we make use of the experimental design of Krott et al. (2001, also chapter 2). Thereafter, we present simulation studies in which we predict the responses of the participants in our experiments with a computational model of analogy, TiMBL, developed by Daelemans, Zavrel, Van der Sloot, & Van den Bosch (2000). With the means of this model, we can simulate the paradigmatic effect of the left and right constituent family. In addition, we can also test whether features of the left constituent, such as rime, gender and inflectional class, affect the selection of linking elements. The latter allows us to test the effect of general rules, like the ones listed in Dressler et al. (2001). In the general discussion, we outline how effects of the constituent family as well as effects of characteristics of the left constituent such as rime or inflectional class can be modeled in a symbolic interactive activation model for analogy.

Experiment 1: the linking -s-

Method

Materials. As in experiments 1 and 2 of Krott et al. (2001, also chapter 2), we constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3). Each set contained 20 nouns, except for L2, for which we could only find 10 nouns. The constituents of L1 and R1 had constituent families with as strong a bias as possible towards the linking element *-s-*. Conversely, L3 and R3 showed a bias as strong as possible against *-s-*. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against *-s-*. We used the CELEX lexical database (Baayen et al., 1995) to determine the constituent families of the constituents in these six sets.

The constituents in the L1 set had constituent family members all of which contained the linking element *-s-*. The mean number of compounds in these fami-

lies was 12.1 (range 5–46). Their mean token frequency was 2417 per 6 million wordforms (range 1–11047). The constituents in the R1 set had constituent family members of all which also contained the linking element *-s-*. The mean number of compounds in these families was 2.3 (range 2–4). Their mean token frequency was 18.1 per 6 million wordforms (range 0–75). The neutral set L2 included left constituents whose families contained between 30% and 70% compounds with the linking element *-s-*. These families had a mean number of compounds of 3.3 (range 2–6) and a mean token frequency of 41.4 per 6 million wordforms (range 0–190). The constituents in the R2 set had constituent family members of which 40% to 60% contained the linking element *-s-*. These families had a mean number of compounds of 5.5 (range 3–15) and a mean token frequency of 69.1 per 6 million wordforms (range 7–437). The remaining sets L3 and R3, the groups with a bias against *-s-*, contained constituents whose family members tend not to occur with the linking *-s-* (L3: 0%; R3: less than 20%). There were in the mean 2.1 (L3: range 1–9) and 2.6 (R3: range 2–6) family members, respectively, with *-s-*. Their mean token frequency was 362.8 (range 0–3490; L3) and 24.3 (range 0–77; R3). These are the maximal contrasts that allowed us to select 20 constituents for each experimental set, except for L2 for which we were able to select 10 nouns.

As in experiments 1 and 2 in Krott et al. (2001, also chapter 2), each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: bias in the left position (positive, neutral, and negative) and bias in the right position (positive, neutral, and negative). None of these compounds is attested in the CELEX lexical database. All have a high degree of semantic interpretability. Appendix A lists all 6 x 20 + 3 x 10 = 150 experimental items.

Procedure. As in Krott et al. (2001, also chapter 2), the participants performed a cloze-task. The experimental list of items was presented to the participants in written form. Each line presented two compound constituents separated by two underscores. We asked the participants to combine these constituents into new compounds and to specify the most appropriate linking element, if any, at the position of the underscores, using their first intuitions. As already mentioned, the first constituent may change its form when it is combined with a linking element. The instructions made clear that these changes were not of interest and could be ignored. Each participant saw the list of items together with the items of the other two experiments presented in this paper in a separate randomized order. The experiment

Table 6.1: Mean number of selected linking elements (maximum = 33) when varying the bias for *-s-* (positive, neutral, and negative) in the left and right compound position. Standard deviations between parentheses.

left position		right position: positive		neutral		negative	
positive	s	30.7	(3.8)	31.4	(2.6)	31.5	(3.1)
	not s	2.3	(3.8)	1.6	(2.6)	1.6	(3.1)
neutral	s	23.5	(7.5)	23.3	(9.5)	24.5	(7.5)
	not s	9.5	(7.5)	9.7	(9.5)	8.5	(7.5)
negative	s	12.0	(6.9)	13.8	(7.8)	14.7	(8.3)
	not s	21.0	(6.9)	19.3	(7.8)	18.3	(8.3)

lasted approximately 25 minutes.

Participants. Thirty-three participants of an introductionary linguistics course at the University of Vienna volunteered to take part in the experiment. All were native speakers of German.

Results and discussion

The participants always filled in a possible German linking element. Therefore, no error was attested. Table 6.1 summarizes the mean number of *s* responses versus other responses for the nine experimental conditions. Appendix A lists the individual words together with the absolute numbers of *s* and *not s* responses.

A by-item logit analysis (see, e.g., Rietveld & Van Hout, 1993) of the *s* and *not s* responses revealed only a main effect of the bias in the left position ($F(2,141) = 64.5$, $p < .001$). There is neither a main effect of the bias in the right position ($F(2,141) < 1$, $p = .439$) nor an interaction of the biases in both positions ($F(4,141) < 1$, $p = .987$). The upper panel of Figure 6.1 shows the large effect of the left bias on the mean number of *s* responses, averaged over items. Surprisingly, the small, but significant effect of the right constituent family that was attested for both Dutch linking elements *-en-* and *-s-* in two different sets of experimental items (Krott et al., 2001, also chapter 2; Krott, Krebbers, Schreuder, & Baayen, in press, also chapter 4) is absent in this experimental set.

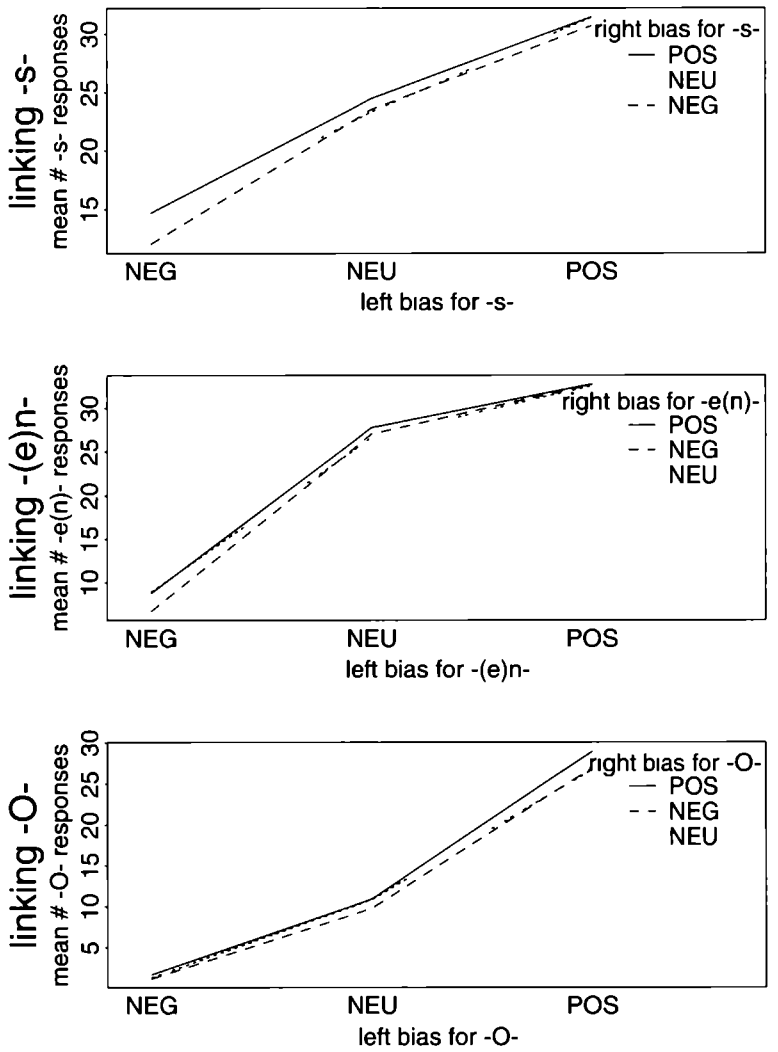


Figure 6.1: Results of the experiments: interaction of biases in the left and right compound position for the linking *-s-* (upper panel), the linking *-(e)n-* (middle panel), and the linking *-∅-* (lower panel). POS: positive bias; NEU: neutral bias; NEG; negative bias.

Experiment 2: the linking -(e)n-

Having tested the effect of the left and right constituent families on the linking *-s-*, we now turn to their effect on the linking elements *-n-* and *-en-*.

Method

Materials. As in experiment 1, we constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3). Each set contained 20 nouns, except for R1, for which we could only find 18 nouns. The constituents of L1 and R1 had constituent families with as strong a bias as possible towards the linking elements *-n-* or *-en-*. Conversely, L3 and R3 showed a bias as strong as possible against *-(e)n-*. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against *-(e)n-*. We used the CELEX lexical database (Baayen et al., 1995) to determine the constituent families of the constituents in these six sets.

The constituents in the L1 set had constituent family members all of which contained the linking element *-(e)n-*. The mean number of compounds in these families was 8.8 (range 5–22). Their mean token frequency was 927.3 per 6 million wordforms (range 0–15066). The constituents in the R1 set had constituent family members of which at least 75% contained the linking element *-(e)n-*. The mean number of compounds in these families was 2.3 (range 2–4). Their mean token frequency was 9.1 per 6 million wordforms (range 0–48). The neutral set L2 included left constituents whose families contained between 40% and 70% compounds with the linking element *-(e)n-*. These families had a mean number of compounds of 2.8 (range 2–6) and a mean token frequency of 89.0 per 6 million wordforms (range 0–707). The constituents in the R2 set had constituent family members of which 40% to 60% contained the linking element *-(e)n-*. These families had a mean number of compounds of 2.7 (range 2–7) and a mean token frequency of 12.3 per 6 million wordforms (range 0–55). The remaining sets L3 and R3, the groups with a bias against *-(e)n-*, contained constituents whose family members tend not to occur with the linking *-(e)n-* (L3: less than 5%; R3: less than 15%). There were in the mean 0.1 (L3: range 0–2) and 2.9 (R3: range 2–6) family members with *-(e)n-* respectively. Their mean token frequency was 2.7 (range 0–54; L3) and 17.3 (range 0–60; R3). These are the maximal contrasts that allowed us to select 20 constituents for each experimental set, except for R3 for which we were able to select 18 nouns.

As in experiment 1, each of the three sets of left constituents (L1, L2, L3) was

Table 6 2 Mean number of selected linking elements when varying the bias for *-(e)n-* (positive, neutral, and negative) in the left and right compound position Standard deviations between parentheses

left position		right position: positive		neutral		negative	
positive	en	32 7	(0 7)	32 5	(0 8)	32 7	(0 5)
	not en	0 4	(0 7)	0 5	(0 8)	0 3	(0 5)
neutral	en	27 1	(7 5)	26 7	(7 7)	27 8	(8 4)
	not en	6 0	(7 5)	6 3	(7 7)	5 2	(8 4)
negative	en	6 8	(7 0)	9 0	(8 8)	8 8	(8 9)
	not en	26 3	(7 0)	24 1	(8 8)	24 2	(8 9)

combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors bias in the left position (positive, neutral, and negative) and bias in the right position (positive, neutral, and negative) None of these compounds is attested in the CELEX lexical database All have a high degree of semantic interpretability Appendix B lists all 6 x 20 + 3 x 18 = 174 experimental items

Procedure The procedure was identical to that of Experiment 1

Participants The participants were identical to those of Experiment 1

Results and discussion

Only one response was unclear and had to be counted as an error All other responses were taken into the analysis Table 6 2 summarizes the mean number of *(e)n* responses versus other responses for the nine experimental conditions Appendix B lists the individual words together with the absolute numbers of *(e)n* and *not (e)n* responses

As in Experiment 1, a by-item logit analysis of the *(e)n* and *not (e)n* responses revealed only a main effect of the bias in the left position ($F(2,165) = 89\,5$, $p < 001$) There is neither a main effect of the bias in the right position ($F(2,165) < 1$, $p = 667$) nor an interaction of the biases in both positions ($F(4,165) < 1$, $p = 937$) This is also visible in the middle panel of Figure 6 1 which shows the effect of the biases on the mean number of *(e)n* responses, averaged over items Apparently, the effect of the bias in the right constituent family is generally absent in German compounds

Experiment 3: the linking possibility -∅-

In this section, we test whether the analogy to the left constituent family is also effective for compounds without a linking element (-∅-). As already mentioned, the -∅- is the default linking possibility in German compounds. Given the recent discussion about morphological defaults (Marcus, Brinkman, Clahsen, Wiese, & Pinker, 1995; Clahsen, 1999), one would expect that default linking elements are governed by rules, not by analogy. However, if the linking elements *-s-* and *-(e)n-* are selected by analogy to their constituent families, the same might be true for -∅-.

Method

Materials. As in experiments 1 and 2, we constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3). Each set contained 20 nouns. The constituents of L1 and R1 had constituent families with as strong a bias as possible towards the linking -∅-. Conversely, L3 and R3 showed a bias as strong as possible against -∅-. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against -∅-. We used the CELEX lexical database (Baayen et al., 1995) to determine the constituent families of the constituents in these six sets.

The constituents in the L1 set had constituent family members all of which contained the linking element -∅-. The mean number of compounds in these families was 15.9 (range 10–28). Their mean token frequency was 1471.4 per 6 million wordforms (range 35–9622). The constituents in the R1 set also had constituent family members of all which contained the linking element -∅-. The mean number of compounds in these families was 7 (range 5–16). Their mean token frequency was 118.7 per 6 million wordforms (range 13–911). Neutral left constituents are rare. The neutral set L2 included left constituents whose families contained between 30% and 70% compounds with the linking element -∅-. These families had a mean number of compounds of 3.3 (range 3–6) and a mean token frequency of 8757.6 per 6 million wordforms (range 0–12203). The constituents in the R2 set had constituent family members of which 30% to 70% contained the linking element -∅-. These families had a mean number of compounds of 7.6 (range 5–15) and a mean token frequency of 104.4 per 6 million wordforms (range 13–579). The remaining sets L3 and R3, the groups with a bias against -∅-, contained constituents whose family members tend not to occur with the linking -∅- (L3: less than 15%; R3: less than 20%). There were in the mean 0.4 (L3: range 0–4) and 0.1 (R3: range 0–1) family members with -∅- respectively. Their mean token frequency was 146.9

Table 6.3: Mean number of selected linking elements when varying the bias for -∅- (positive, neutral, and negative) in the left and right compound position. Standard deviations between parentheses.

		right position					
left position		positive		neutral		negative	
positive	∅	26.9	(4.6)	26.7	(7.6)	29.0	(4.0)
	not ∅	6.1	(4.6)	6.3	(7.6)	4.1	(4.0)
neutral	∅	9.8	(9.2)	10.9	(10.0)	10.9	(10.0)
	not ∅	23.3	(9.2)	22.2	(10.0)	22.1	(10.0)
negative	∅	1.2	(3.5)	1.3	(3.2)	1.7	(4.0)
	not ∅	31.8	(3.5)	31.7	(3.2)	30.3	(4.0)

(range 0–1757; L3) and 0.4 (range 0–4; R3). These are the maximal contrasts that allowed us to select 20 constituents for each experimental sets.

As in experiments 1 and 2, each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: bias in the left position (positive, neutral, and negative) and bias in the right position (positive, neutral, and negative). None of these compounds is attested in the CELEX lexical database. All have a high degree of semantic interpretability. Appendix B lists all 9 x 20 = 180 experimental items.

Procedure. The procedure was identical to that of Experiments 1 and 2.

Participants. The participants were identical to those of Experiments 1 and 2.

Results and discussion

Only once a participant responded with a letter that never occurs as a linking element. This response was counted as an error. Table 6.3 summarizes the mean number of -∅- responses versus other responses for the nine experimental conditions. Appendix C lists the individual words together with the absolute numbers of -∅- and *not* -∅- responses.

A by-item logit analysis of the -∅- and *not* -∅- responses again revealed only a main effect of the bias in the left position ($F(2,171) = 226.7$, $p < .001$). There is neither a main effect of the bias in the right position ($F(2,171) < 1$, $p = .595$) nor an interaction of the biases in both positions ($F(4,171) < 1$, $p = .953$). The lower panel of Figure 6.1 shows the effect of the biases on the mean number of -∅- responses.

These results confirm the hypothesis that the right constituent family does not affect German linking elements.

Our hypothesis that the analogical effect of the left constituent family is not only relevant for *-(e)n-* and *-s-*, but also for the *-∅-* has been confirmed. Thus, even the default compounding formation is, at least in part, analogically determined.

A comparison of the results of the three experiments shows that a neutral left bias for the *-∅-* leads to fewer *-∅-* responses (10.5) than a neutral bias for *-s-* or *-(e)n-* leads to *s* (23.8) or *(e)n* (27.1) responses, respectively. These differences are significant (*-s-* versus *-∅-*: $t_2(88) = 6.5$; $p < .001$; *-(e)n-* versus *-∅-*: $t_2(116) = 10.4$; $p < .001$), while the number of responses for *-s-* and *-(e)n-* differ only marginally from each other ($t_2(86) = 1.9$; $p = .058$). The reduced number of *-∅-* responses cannot be due to different strengths of the biases in the different experiments, because these were very similar (mean bias for *-s-*: 53.5; mean bias for *-(e)n-*: 56.0; mean bias for *-∅-*: 52.9). Interestingly, this result is in line with an earlier finding for Dutch linking elements. Krott et al. (in press-a, also chapter 3) report that a bias for *-∅-* can be violated in Dutch compounds more easily than a bias for *-en-* or *-s-*. Thus, although using no linking element is the most common way of forming compounds in both languages, they share the tendency for using overt linking elements.

Modeling German linking elements

In Krott et al. (2001, also chapter 2) we have shown that selected linking elements for novel Dutch compounds, as they are given by the participants in production experiments, can be modeled with a high degree of accuracy using an exemplar-based machine-learning algorithm for the modeling of analogy, TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2000). Exemplar-based learning models combine similarity-based reasoning with the extensive storage of exemplars in an instance database. The class of a target, i.e. its outcome, is determined by comparing the target with the exemplars in the instance base using a set of user-specified features.[1] The most similar instance or the set of the most similar instances is used as the prediction basis.

The simulation studies of Krott et al. revealed that the crucial analogical factor for predicting Dutch linking elements is the left constituent, which represents the left constituent family. Prediction accuracy was enhanced when semantic class infor-

[1]For a description of the model's similarity metrics, see Daelemans et al. (2000) and Krott et al. (2001, also chapter 2).

mation of the right constituent was included in the feature set. Addition of the second constituent to the set did not improve prediction accuracy, although production experiments revealed clear evidence for the existence of an analogical effect of the second constituent, a non-semantic effect (Krott, Krebbers, Schreuder, & Baayen, in press, also chapter 4).

The question arises whether the choice of the linking elements in German novel compounds can also be predicted with an exemplar-based modeling technique. Is it again the paradigmatic set of the left constituent family that leads to higher prediction accuracies? Dressler, Libben, Stark, Pons & Jarema (2001) report that German linking elements are selected on the basis of ten categories of left constituents, which they interpret as evidence for rules. However, they also mention some evidence suggesting a role for analogical effects of constituent families. Simulation studies with TiMBL allow us to test whether the selected linking elements can be predicted more accurately on the basis of the left constituent family or on the basis of properties of the left constituent such as phonology, gender, and inflectional class.

As a baseline study, we first ascertain to what extent constituent families and properties of the left constituent predict the linking elements of existing German compounds, namely the 8331 German compounds listed in CELEX. Table 6.4 lists the features that we have investigated, namely the left constituent (C1), the right constituent (C2), and rime, gender, and inflectional class of the left constituent. TiMBL provides for each feature a relevance weight, the information gain (IG). The information gain measures how much information the feature contributes to the classification process. It therefore provides a first estimation of the prediction relevance of a feature. The column labeled 'celex' of Table 6.4 lists the information gain values for the selected features, when TiMBL is trained on all 8331 compounds in CELEX. The left constituent, and therefore the left constituent family, has the highest information gain value (1.73), followed by the rime of the left constituent (1.06) and the right constituent (.86). Less relevant for the classification are the inflectional class (0.2) and the gender (0.4) of the left constituent.

These values differ from the values obtained in the training for the production experiments (-S-, -EN-, and -∅-). This difference arises due to different classification procedures. While linking elements in the CELEX compounds were classified as *-s-*, *-(e)n-*, *-∅-* etc., they were classified as either *-s-* or *not -s-* in the *-s-* experiment and as either *-(e)n-* or *not -(e)n-* in the *-(e)n-* experiment. In all experiments, just as in the baseline study, the left constituent reveals the highest information gain

Table 6.4: Feature sets used in the TiMBL simulations studies of all German compounds in CELEX (celex) and the three experiments (-S-, -EN-, -∅-) as well as their Information Gain. C1: left constituent; C2: right constituent; rime: rime of C1; gender: gender of C1; inflection: inflectional class of C1.

		Experiments		
features	celex	-S-	-EN-	-∅-
C1	1.73	.64	.56	.83
C2	.86	.27	.22	.31
rime	1.06	.35	.32	.36
gender	.24	.02	.09	.08
inflection	.52	.04	.23	.18

Table 6.5: Feature sets used in the TiMBL simulations studies of all German compounds in CELEX (celex) and the three experiments (-S-, -EN-, -∅-) and their prediction accuracies in percentage of correctly predicted linking elements. C1: left constituent; C2: right constituent; rime: rime of C1; gender: gender of C1; inflection: inflectional class of C1.

		Experiments		
features	celex	-S-	-EN-	-∅-
C1	87.4	79.3	79.9	80.6
C1,C2	86.9	79.3	79.9	80.6
rime	79.0	50.0	82.8	76.7
rime,gender,inflection	84.0	62.0	88.5	82.2
C1,rime,gender,inflection	91.9	79.3	79.9	80.6
agreement among participants		81.8	89.1	87.4

value. In contrast to the baseline study, the experiments suggest that the right constituent is the second most relevant feature. A comparison of the three experiments shows that gender is more important in the *-(e)n-* experiment than in the other experiments, while the inflectional class is more important in the -∅- experiment. The feature rime is most relevant in the *-s-* experiment. On the basis of these values, we expect that the left constituent will be the strongest predictor of German linking elements in novel compounds, followed by the right constituent. The remaining features are expected to be more or less relevant depending on the set of target compounds.

Table 6.5 lists the percentage of correctly predicted linking elements in the exist-

ing German compounds in CELEX as well as in the production experiments.[2] The prediction accuracies given in the column 'celex' are obtained by a 'leave-one-out' procedure in which each CELEX compound is predicted on the remaining compounds. The highest prediction accuracy for a single feature is obtained by using the left constituent (87.4%). This has also been the case for the prediction of linking elements in existing Dutch compounds, although there, the left constituent predicts the selection somewhat better (92.6%) (Krott et al., in press-a, also chapter 3). Note that in both languages, the model did not simply select the most frequent linking possibility. Otherwise, it would have reached a prediction accuracy of only 65%, which is the percentage of German compounds that do not contain any linking element. Surprisingly, including the right constituent in the training, the feature with the second highest information gain value, does not lead to an increase, but to a slight decrease in prediction accuracy (86.9%) of German linking elements. However, this result is in line with the results of the production experiments, in which the right constituent also did not affect the selection of linking elements. The combination of characteristics of the left constituent, i.e. rime, gender, and inflectional class, reaches a prediction accuracy of 84%, which is significantly lower than the prediction reached by the left constituent (proportions test: $p < .001$). However, taking left constituent and its properties together leads to the high accuracy score of 91.9%, which is significantly higher than that obtained on the basis of the left constituent by itself (proportions test: $p < .001$). Similarly, in the case of Dutch compounds, the combination of the left constituent and the rime and the suffix of the left constituent led to a higher prediction accuracy (93.4%) than the left constituent by itself (Krott et al., in press-a, also chapter 3). Thus, neither the left constituent nor its characteristics alone are sufficient to predict linking elements in existing German noun-noun compounds. It appears to be that both factors are relevant simultaneously, albeit with different weights.

The simulation studies of the responses given for novel compounds in the production experiments, however, reveal a somewhat different pattern of results. In order to predict the choices in the experiments, we compared the TiMBL's predictions with the selected linking elements that were chosen by the majority of the participants. As Table 6.5 shows, in both the *-s-* and the *-∅-* experiment, the majority choices are most accurately predicted by the left constituent (*-s-*: 79.3%;

[2]For all reported prediction accuracies, the following parameter settings were used: similarity algorithm: IB1; feature metrics = weighted overlap, features weighed by information gain values, size of best neighbor set = 1. Different settings do not change the pattern of results. For detailed information about the parameters, see Daelemans et al. (2000).

-∅-: 80.6%). Including the right constituent in the feature set does not change the results. Using just the characteristics of the left constituent leads to a decrease in prediction accuracy in the *-s-* experiment (62.0%; proportions test: $p = .002$), while it leads to a slight increase in prediction accuracy in the *-∅-* experiment (82.2%), which is, however, not significant (proportions test: $p = .787$). Surprisingly, in contrast to the baseline study, the combination of the left constituent and its characteristics does not improve the prediction accuracy. A different pattern emerges for the *-(e)n-* experiment. Here, combining the left constituent and its characteristics also does not increase the prediction accuracy obtained by the left constituent alone (79.9%; trained on the constituent families of the experiment). However, gender, rime, and inflectional class of the left constituent reveal a significantly higher prediction accuracy (88.5%; proportions test: $p = .040$). This result is mainly due to the rime, which alone already correctly predicts 82.8%.

Summing up, in the case of existing German compounds, a combination of the left constituent and its characteristics leads to the highest prediction accuracy. In the case of the *-s-* experiment, responses were predicted quite well by just the left constituent. In the *-(e)n-* experiment, responses are better predicted by the set of gender, rime, and inflectional class. In the *-∅-* experiment, the left constituent and the set of its properties led to very similar prediction accuracies.

One might argue that the training set of 8331 German compounds is somewhat small, when compared to the 32,000 compounds in the Dutch simulation studies. We therefore included 24,000 German compounds into the training set that were extracted out of two German newspaper corpora, Frankfurter Rundschau and Stuttgarter Zeitung, which contain 76 million wordforms when combined. This allowed us to examine the effect of the two constituent families in a much broader database. This increase of training data leads to a significantly higher prediction accuracy when predicting the existing compounds in CELEX on the basis of the left constituent (93.4% versus 87.4%; proportions test: $p < .001$). However, the prediction accuracies obtained with the left constituent changed only marginally for the novel compounds used in our experiments (*-s-*: 80.0%, $p = 1$; *-(e)n-*: 78.2%, $p = .792$; *-∅-*: 81.7%, $p = .893$; proportions tests). As in all previous simulation studies, the right constituent did not contribute to the prediction accuracy at all. We conclude that the prediction of the sometimes idiosyncratic patterns of linking elements in existing compounds can be improved by extending the training set. However, the patterns that are relevant for predicting linking elements in novel compound are already captured by the small set of the CELEX compounds.

The bottom row of Table 6.5 lists for all three experiments the mean percentages of participants that chose the linking elements that were selected by the majority of the participants. In the case of the *-s-* experiment, in the mean, 81.1% of the participants agreed with the majority choice for a linking element, while the highest prediction accuracy, based on the left constituent, was 79.3%. In the *-(e)n-* experiment, 89.1% of the participants agreed with the majority choice, while the model reaches a prediction accuracy of 88.5%, if the training is based on the rime, the gender, and the inflectional class of the left constituent. The difference between the participants' agreement (87.4%) and the model's prediction (79.4%; training on left constituent) in the *-∅-* experiment is not significant (proportion test: $p = .115$). We therefore conclude that, taking the highest prediction accuracies for each experiment, participants and the model appear to find the task equally difficult in all experiments. The same result was found in the simulation studies of Dutch compounds in Krott et al. (2001, also chapter 2).

We conclude that the left constituent is the strongest predictor of linking elements in German noun-noun compounds. However, depending on the class of the left constituent, characteristics such as gender, inflectional class, and, in particular, the rime either enhance the prediction or lead to a better prediction than the constituent itself. Apparently, these factors all play a role. However, their relevance seems to vary somewhat with the type of the left constituent.

General discussion

In this study, we focused on the paradigmatic analogical effect of the constituent families on the selection of linking elements in novel German compounds. We conducted three production experiments in which participants had to select the appropriate linking element for novel compounds, and explained the choices of the participants with an exemplar-based computational model for analogy, TiMBL (Daelemans et al., 2000).

In all three production experiments, we observed a strong paradigmatic effect of the left constituent family on the selection of linking elements, just as reported in previous studies for Dutch linking elements (Krott et al., 2001, also chapter 2; Krott, Krebbers, Schreuder, & Baayen, in press, also chapter 4). A strong bias for a particular linking element in the left constituent family leads to more responses with this linking element. We could not, however, replicate the small, but significant paradigmatic effect of the right constituent family that has been found for Dutch

linking elements. The choice of German linking elements appears to be made on the basis of proportions of the left constituents only.

A comparison of the three experiments revealed that a left positive bias for *-∅-* is less effective than a left positive bias for *-s-* or *-(e)n-*. A left bias for *-∅-* is more easily overruled, a finding that has also been attested for Dutch linking elements (Krott et al., in press-a, also chapter 3). This is surprising considering the fact that the *-∅-* is the default linking element in both German and Dutch compounds.

Simulation studies with the exemplar-based model TiMBL, addressing both the prediction of linking elements in existing compounds and in novel compounds presented in the experiments, confirmed that the left constituent is the strongest predictor of linking elements in German noun-noun compounds. Just as in the experiments, the right constituent family does not contribute to a higher prediction accuracy. In the case of the *-s-* experiment, the left constituent family is the analogical factor with the highest independent prediction accuracy, which cannot be enhanced any further by including other factors. However, by adding gender, inflectional class, and, in particular, the rime of the left constituent to the feature set, we can improve the prediction accuracies for existing compounds. The combination of rime, gender, and inflectional class (without left constituent) leads to the highest prediction accuracy in the case of the *-(e)n-* experiment. We therefore conclude that it is neither the constituent family by itself nor properties such as rime, gender, and the inflectional class that affect the choice of linking elements, but an interplay of these factors.

Although we did not include the categories of linking elements identified by Dressler et al. (2001) in our experimental design, a post-hoc analysis shows that each experiment represents predominantly a particular subset of Dressler et al.'s categories. The sets of items with a positive and neutral bias for *-s-* in Experiment 1 mainly contain nouns of Dressler et al.'s categories 6 and 7, i.e. sets that both prefer the linking *-s-*. The set with a negative bias for *-s-* mainly contains items of categories 3 and 4, nouns that are typically combined with *-n-* and *-en-*. In the case of the *-(e)n-* experiment, all three sets mainly contain nouns of categories 2 and 4, i.e. nouns that are typically combined with *-n-* and *-en-*. Interestingly, 18 out of the 20 left constituents with a negative bias for *-(e)n-* belong to categories that, according to Dressler et al., should be combined with *-(e)n-*. In the production experiment, however, only 24% of these items were responded to with *-(e)n-*. In these cases, the constituent family clearly emerges as the stronger force. This is also true for the items in the *-∅-* experiment. These nouns mainly belong to categories that are

combined with *-(e)n-* and *-s-*. Despite the predictions of the categories, participants followed the bias of the constituent families and responded with *-∅-*. For instance, the items with a positive bias for *-∅-* elicited a *-∅-* response in 83.3% of all cases, instead of *-en-* or *-s-*, as predicted by Dressler et al.

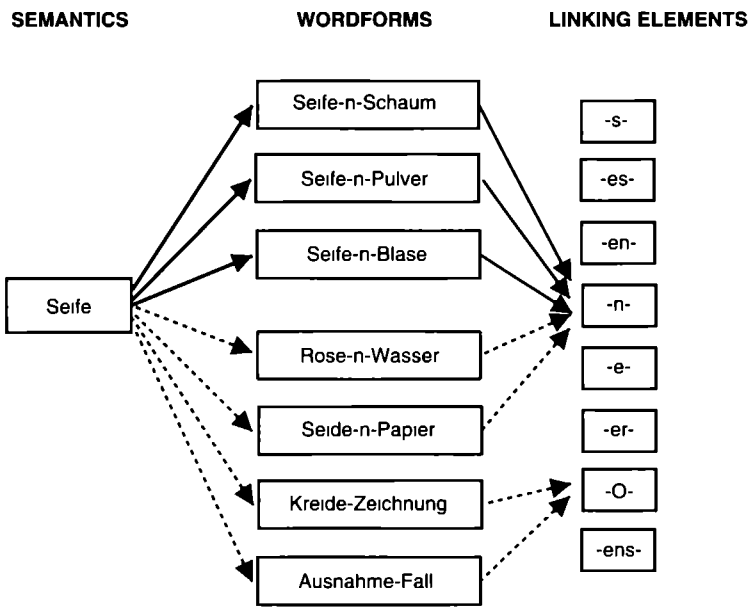


Figure 6.2: Connectivity in a sample part of the lexicon that is involved in the selection of the linking element for the novel German compound *Seife+?+Stift* ('soap pen'). Semantic representations (left layer); wordforms representations (lexemes in the sense of Levelt (1989), central layer) with left constituent family (upper part) and compounds sharing rime, gender, and inflectional class of the first constituent (lower part); linking elements (right layer). Line type represents amount of activation flow (solid arrow: high activation; dotted arrow: low activation).

Considering the combined results of the simulation studies and the production experiments, both in the present study and in the study by Dressler et al., we conclude that, in contrast to Dutch linking elements, German linking elements are chosen on the basis of the left constituent family as well as on the basis of properties of the left constituent such as rime, gender, and inflectional class. The functional role of gender, rime, and inflection class can be construed as evidence for rules that function independently of any stored exemplars, as proposed by Dressler et al. In the approach of this contribution, there are no abstract generalizations. The

effect of properties of the left constituent can be understood as being paradigmatic analogical in nature. This has become evident in the simulation studies with TiMBL.

We can account for these paradigmatic effects in an interactive activation framework, as developed by Krott, Schreuder, & Baayen (in press-b, also chapter 5). They report a computational symbolic interactive activation model that captures the analogical effect of the constituent families on the choice of linking elements in Dutch compounds. In this model, the left and right constituent of a target compound activate the compounds of their constituent families, which in their turn activate their linking elements. The selection of German linking elements can be understood along similar lines. A novel compound can activate both its left constituent family and the constituent families of other left constituents that share features such as rime, gender, and inflectional class. Figure 6.2 illustrates the activation flow for the novel compound *Seife+?+Stift* 'soap pen'. The semantic representation of the left constituent *Seife* sends activation to the members of its constituent family on the wordform level, such as *Seife+n+Schaum* 'lather', *Seife+n+Pulver* 'soap powder', and *Seife+n+Blase* 'soap-bubble'. In addition, it also sends activation to compounds whose left constituent are feminine nouns that end in schwa, such as *Rose+n+Wasser* 'rose water' *Seide+n+Papier* 'tissue paper', *Kreide+Zeichnung* 'chalk drawing', and *Ausnahme+Fall* 'exceptional case'. All these compounds then propagate activation onwards to their linking elements. The linking element that receives the most activation is selected for insertion in *Seife+?+Stift*. In our experiment, it was the *-n* that was chosen most often (94%) for this particular compound. This example shows that, even if the left constituent family has a strong bias for a linking element, *-n-* in our case, compounds sharing the rime can activate other linking elements, such as the *-∅-*, as well. Given that the left constituent was the strongest predictor in our simulation studies, we assume that the left constituent family passes on more activation to the linking elements than compounds whose left constituents are, for instance, feminine nouns that end in schwa. This is represented in Figure 6.2 by different line types of the connections (solid arrows: high activation; dotted arrow: low activation).

The outlined model presupposes that linking elements constitute independent units in the mental lexicon. This allows the model to explain the paradigmatic effects of left constituents sharing a property such as the inflectional class. Independent support for the hypothesis that linking elements are processed as separate units is provided by a visual perception study reported in Dressler et al. (2001). Nevertheless, the strong effect of the left constituent and its properties on the selection

of linking elements reveals a tight connection between the left constituent and the linking element. Note that linking elements are part of the constituent's final syllable and that they group with the left constituent in coordinational structures such as the *-s-* in *Verwaltungs- und Kundendienst* ('administration and customer service'). This tight link between the left constituent and the linking element can be formalized by analyzing the left constituent and its linking element as a compound stem, as proposed by Fuhrhop (1998). We will remain agnostic with respect to the relevance of the notion of the compounding stem and restrict ourselves to observing that, if so required, our psychological model can be understood as the mechanism underlying the creation of compounding stems.

References

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Becker, T.: 1992, Compounding in German, *Rivista di Linguistica* **4**(1), 5–36.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 2000, TiMBL: Tilburg Memory Based Learner Reference Guide. Version 3.0, *Technical Report ILK 00-01*, Computational Linguistics Tilburg University.

Dressler, W. U. and Merlini Barbaresi, L.: 1991, Elements of morphopragmatics, *in* J. Verschueren (ed.), *Levels of Linguistic Adaptation. Selected papers of the International Pragmatics Conference, Antwerp, August 17–22, 1987*, Vol. II, John Benjamins, Amsterdam/Philadelphia, pp. 33–51.

Dressler, W. U., Libben, G., Stark, J., Pons, C. and Jarema, G.: 2001, The processing of interfixed German compounds, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp. 185–220.

Fuhrhop, N.: 1996, Fugenelemente, *in* E. Lang and G. Zifonun (eds), *Deutsch - Typologisch*, de Gruyter, Berlin, pp. 525–550.

Fuhrhop, N.: 1998, *Grenzfälle Morphologischer Einheiten (Border Cases of Morphological Units)*, Stauffenburg, Tuebingen.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Krott, A., Baayen, R. H. and Schreuder, R.: 2001, Analogy in morphology: modeling the choice of linking morphemes in Dutch, *Linguistics* **39**(1), 51–93.

Krott, A., Krebbers, L., Schreuder, R. and Baayen, R. H.: in press, Semantic influence on linkers in Dutch noun-noun compounds, *Folia Linguistica*.

Krott, A., Schreuder, R. and Baayen, R. H.: in press-a, Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds, *in* R. Skousen (ed.), *Analogical Modeling: An Exemplar-Based Approach to Language*.

Krott, A., Schreuder, R. and Baayen, R. H.: in press-b, Linking elements in Dutch noun-noun compounds: constituent families as predictors for response latencies, *Brain and Language*.

Levelt, W.: 1989, *Speaking. From intention to articulation*, The MIT Press, Cambridge, Mass.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R. and Pinker, S.: 1995, German inflection: The exception that proves the rule, *Cognitive Psychology* **29**, 189–256.

Rietveld, T. and Van Hout, R.: 1993, *Statistical Techniques for the Study of Language and Language Behaviour*, Mouton de Gruyter, Berlin.

Unbegaun, B.: 1967, *Russian Grammar*, Clarendon Press.

Van den Toorn, M. C.: 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I), *De Nieuwe Taalgids* **75**(1), 24–33.

Van den Toorn, M. C.: 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for the regulation of linking phonemes in nominal compounds II), *De Nieuwe Taalgids* **75**(2), 153–160.

Appendices

Appendix A

Materials for Experiment 1: left constituent and right constituent (number of *s* responses, number of other responses).

L1-R1: Left Position: Positive *-s-* Bias; Right Position: Positive *-s-* Bias:
Verkehr Ideal (33,0); Handel Möglichkeit (33,0); Unglück Dauer (33,0); Zeitung Verbrechen (32,1); Seemann Bilanz (33,0); Übergang Drang (32,1); Amt Entwicklung (32,1); Versuch Gefährte (33,0); Staat Votum (28,5); Geburt Korrespondent (20,13); Durchschnitt Urkunde (33,0); Volk Formular (33,0); Wolf Manöver (28,5); Weihnacht Verbrecher (33,0); Alter Ausweis (33,0); Leben Körper (33,0); Ort Radius (31,2); Ausgleich Erfahrung (33,0); Teufel Gesuch (32,1); Leiden Koeffizient (31,2)

L1-R2: Left Position: Positive *-s-* Bias; Right Position: Neutral *-s-* Bias:
Volk Zustand (32,1); Durchschnitt Besuch (33,0); Verkehr Vertrag (33,0); Teufel Grad (33,0); Amt Heim (28,5); Übergang Kirche (33,0); Zeitung Beamte (32,1); Alter Summe (33,0); Ort Gesellschaft (30,3); Geburt Form (22,11); Ausgleich Grenze (32,1); Wolf Gruppe (31,2); Handel Wissenschaft (33,0); Seemann Hilfe (32,1); Unglück Leistung (33,0); Weihnacht Apparat (33,0); Staat Kraft (29,4); Versuch Freiheit (32,1); Leiden Bereich (31,2); Leben Zeugnis (33,0)

L1-R3: Left Position: Positive *-s-* Bias; right constituent: Negative *-s-* Bias:
Leben Wechsel (32,1); Teufel Fest (33,0); Staat Buch (23,10); Unglück Preis (33,0); Ausgleich Musik (33,0); Alter Baum (31,2); Zeitung Sucht (32,1); Amt Bruch (29,4); Wolf Wagen (32,1); Übergang Schutz (33,0); Verkehr Industrie (33,0); Handel Karte (32,1); Geburt Bericht (18,15); Versuch Linie (31,2); Seemann Leder (33,0); Weihnacht Versicherung (33,0); Ort Spiel (28,5); Leiden Wand (31,2); Volk Druck (31,2); Durchschnitt Fahrt (33,0)

L2-R1: Left Position: Neutral *-s-* Bias; Right Position: Positive *-s-* Bias:
Schwein Gefährte (17,16); Ausfall Bilanz (31,2); Gut Urkunde (29,4); Verband Formular (31,2); Mitglied Verbrechen (26,7); Himmel Drang (32,1); Kalb Ideal (29,4); Tabak Votum (14,19); Stab Entwicklung (24,9); Mord Möglichkeit (12,21)

L2-R2: Left Position: Neutral *-s-* Bias; Right Position: Neutral *-s-* Bias:
Stab Zustand (25,8); Himmel Freiheit (31,2); Ausfall Summe (31,2); Verband Bereich (32,1); Schwein Heim (9,24); Tabak Apparat (6,27); Gut Hilfe (27,6); Kalb Wissenschaft (29,4); Mord Form (16,17); Mitglied Freiheit (27,6)

L2-R3: Left Position: Neutral *-s-* Bias; Right Position: Negative *-s-* Bias:
Kalb Fahrt (28,5); Schwein Fest (10,23); Mitglied Wand (26,7); Stab Linie (20,13); Himmel Musik (33,0); Tabak Leder (15,18); Verband Versicherung (29,4); Gut Preis (22,11); Ausfall Karte (32,1); Mord Baum (20,13)

L3-R1: Left Position: Negative *-s-* Bias; Right Position: Positive *-s-* Bias:
Abbruch Erfahrung (23,10); Nachricht Moeglichkeit (5,28); Überzeit Gesuch (13, 20); Großmacht Gefährte (24,9); Abfall Bilanz (24,9); Auflauf Koerper (13,20); Auswahl Korrespondent (17,16); Preßluft Dauer (1,32); Unterschrift Formular (12, 21); Antiquität Ausweis (10,23); Absprung Urkunde (33,0); Heimat Ideal (13,20); Gewalt Radius (7,26); Austausch Verbrechen (11,22); Versand Entwicklung (16,17); Seenot Manoever (13,20); Ausruf Votum (30,3); Haftpflicht Verbrecher (6,27); Unlust Drang (12,21); Umwelt Koeffizient (11,22)

L3-R2: Left Position: *-s-* bias; Right Position: Neutral *-s-* Bias:
Haftpflicht Summe (8,25); Heimat Beamte (14,19); Überzeit Grenze (10,23); Preßluft Bereich (1,32); Abbruch Vertrag (26,7); Abfall Apparat (16,17); Seenot Zustand (17,16); Nachricht Wissenschaft (7,26); Austausch Form (14,19); Auflauf Hilfe (9,24); Antiquität Zeugnis (13,20); Großmacht Freiheit (23,10); Auswahl Kirche (20,13); Versand Leistung (14,19); Gewalt Besuch (10,23); Unlust Gesellschaft (10,23); Umwelt Heim (1,32); Absprung Grad (25,8); Ausruf Kraft (29,4); Unterschrift Gruppe (8,25)

L3-R3: Left Position: Negative *-s-* Bias; Right Position: Negative *-s-* Bias:
Nachricht Buch (2,31); Antiquität Schutz (7,26); Unterschrift Baum (12,21); Heimat Bruch (13,20); Großmacht Wechsel (17,16); Absprung Wagen (19,14); Preßluft Musik (2,31); Versand Leder (17,16); Abfall Versicherung (18,15); Überzeit Linie (7,26); Haftpflicht Spiel (5,28); Gewalt Industrie (6,27); Unlust Sucht (8,25); Austausch Fahrt (13,20); Umwelt Druck (2,31); Abbruch Bericht (18,15); Auswahl Wand (20,13); Seenot Karte (15,18); Ausruf Preis (26,7); Auflauf Fest (13,20)

Appendix B

Materials for Experiment 2: left constituent and right constituent (number of *-(e)n-* responses, number of other responses).

L1-R1: Left Position: Positive *-(e)n-* Bias; Right Position: Positive *-(e)n-* Bias: Rose Reiter (32,1); Kette Staub (32,1); Bär Deck (33,0); Küche Lärm (33,0); Straße Rauch (33,0); Suppe Honig (33,0); Zitrone Angebot (33,0); Seite Last (33,0); Börse Heft (32,1); Seife Strauß (32,1); Hölle König (33,0); Schütze Haß (33,0); Tasche Jäger (33,0); Stange Wärter (32,1); Tanne Nest (33,0); Woche Schmaus (33,0); Tinte Kugel (33,0); Glocke Batterie (33,0)

L1-R2: Left Position: Positive *-(e)n-* Bias; Right Position: Neutral *-(e)n-* Bias: Börse Reihe (31,2); Nerv Gewebe (32,1); Küche Leben (33,0); Tinte Löffel (33,0); Seife Stift (31,2); Tanne Gebirge (33,0); Treppe Bett (33,0); Stange Material (33,0); Glocke Bier (33,0); Bär Hals (32,1); Schütze Gesang (33,0); Straße Schein (33,0); Seite Zaun (33,0); Rose Zimmer (33,0); Kette Hieb (33,0); Hölle Wald (31,2); Suppe Archiv (32,1); Woche Vater (32,1); Zitrone Salat (33,0); Tasche Spitze (33,0)

L1-R3: Left Position: Positive *-(e)n-* Bias; right constituent: Negative *-(e)n-* Bias: Zitrone Ball (33,0); Bär Tag (33,0); Tinte Zeichen (33,0); Glocke Bruch (33,0); Stange Stück (33,0); Straße Land (31,2); Seite Tuch (33,0); Küche Straße (32,1); Kette Arbeit (33,0); Rose Bank (33,0); Seife Blume (33,0); Suppe Meister (33,0); Schütze Karte (33,0); Tasche Schiff (33,0); Treppe Weg (33,0); Tanne Papier (32,1); Woche Zeit (33,0); Hölle Recht (33,0); Nerv Bild (32,1); Börse Geschichte (31,2)

L2-R1: Left Position: Neutral *-(e)n-* Bias; Right Position: Positive *-(e)n-* Bias: Herr Deck (31,2); Schanze Lärm (31,2); Sinn Strauß (1,32); Christ Reiter (15,18); Alp König (33,0); Kohle Wärter (30,3); Fels Staub (22,11); Ehre Schmaus (33,0); Aufgabe Haß (32,1); Asche Kugel (30,3); Weide Jäger (27,6); Sekunde Rauch (33,0); Schwester Heft (33,0); Rebe Last (33,0); Ohr Batterie (28,5); Eiche Nest (33,0); Schmiere Honig (22,11); Scheibe Angebot (33,0)

L2-R2: Left Position: Neutral *-(e)n-* Bias; Right Position: Neutral *-(e)n-* Bias: Schwester Gebirge (25,8); Schmiere Stift (18,14); Aufgabe Archiv (31,2); Rebe Zaun (31,2); Eiche Spitze (33,0); Ehre Vater (32,1); Sinn Schein (1,32); Asche Bett (28,5); Sekunde Leben (33,0); Ohr Hals (27,6); Fels Bier (28,5); Kohle Ge-

webe (28,5); Irre Wald (21,12); Herr Löffel (30,3); Weide Reihe (25,8); Scheibe Salat (32,1); Christ Gesang (17,16); Schanze Material (30,3); Achse Hieb (33,0); Alp Zimmer (30,3)

L2-R3: Left Position: Neutral *-(e)n-* Bias; Right Position: Negative-*(e)n-* Bias: Schwester Tag (30,3); Asche Ball (30,3); Kohle Straße (31,2); Aufgabe Bild (31,2); Ehre Meister (32,1); Herr Bank (32,1); Alp Zeichen (26,7); Schmiere Tuch (21,12); Sekunde Arbeit (33,0); Sinn Zeit (2,31); Scheibe Blume (33,0); Weide Recht (23,10); Christ Geschichte (21,12); Eiche Papier (32,1); Achse Stück (33,0); Schanze Karte (33,0); Irre Land (22,11); Ohr Bruch (21,12); Fels Schiff (24,9); Rebe Weg (31,2)

L3-R1: Left Position: Negative *-(e)n-* Bias; Right Position: Positive *-(e)n-* Bias: Kreide Rauch (16,17); Welt Schmaus (21,12); Bank König (26,7); Flut Jäger (23,10); Saat Angebot (8,25); Aktion Lärm (0,33); Sensation Reiter (1,32); Kultur Last (7,26); Industrie Deck (1,32); Staat Haß (8,25); Granat Staub (16,17); Aufsicht Wärter (0,33); Schicht Honig (17,16); Hochzeit Heft (0,33); Tür Kugel (10,23); Ansicht Batterie (2,31); Zeitung Nest (0,33); Fabrik Strauß (3,30)

L3-R2: Left Position: *-(e)n-* bias; Right Position: Neutral *-(e)n-* Bias: Bank Hieb (20,13); Ansicht Material (1,32); Arznei Salat (12,21); Fabrik Leben (2,31); Sensation Wald (1,32); Schicht Gewebe (15,18); Kreide Hals (24,9); Granat Schein (13,20); Aufsicht Reihe (0,33); Saat Löffel (7,26); Tür Spitze (9,24); Staat Bier (9,24); Aktion Bett (0,33); Industrie Gebirge (1,32); Kultur Gesang (3,30); Zeitung Stift (0,33); Partei Zimmer (21,12); Hochzeit Archiv (0,33); Welt Vater (26,7); Flut Zaun (15,18)

L3-R3: Left Position: Negative *-(e)n-* Bias; Right Position: Negative *-(e)n-* Bias: Aktion Land (0,33); Tür Zeichen (5,28); Schicht Papier (12,21); Arznei Tuch (9,24); Ansicht Straße (0,33); Granat Bild (8,25); Staat Arbeit (4,29); Hochzeit Geschichte (1,32); Zeitung Bank (0,33); Welt Recht (21,12); Industrie Ball (2,31); Partei Schiff (21,12); Fabrik Tag (3,30); Flut Weg (11,22); Kultur Karte (3,30); Aufsicht Zeit (1,32); Kreide Bruch (18,15); Saat Blume (4,29); Bank Stück (12,21); Sensation Meister (0,33)

Appendix C

Materials for Experiment 3: left constituent and right constituent (number of -ø- responses, number of other responses).

L1-R1: Left Position: Positive -ø- Bias; Right Position: Positive -ø- Bias:
Wald Bombe (29,4); Tür Eisen (28,5); Berg Läufer (31,2); Preis Säule (33,0); Zahn Gerät (32,1); Zug Gelenk (15,18); Stein Schrift (29,4); Rohr Meter (28,5); Atom Wolle (33,0); Stadt Nummer (26,7); Tier Monat (29,4); Wand Note (32,1); Herz Analyse (28,5); Fisch Wurst (27,6); Transport Wächter (26,7); Tisch Kern (29,4); Mond Wolke (30,3); Tee Flasche (33,0); Fest Beere (28,5); Öl Essen (33,0)

L1-R2: Left Position: Positive -ø- Bias; Right Position: Neutral -ø- Bias:
Mond Boot (28,5); Wand Bett (31,2); Stein Spiegel (28,5); Wald Sprache (27,6); Öl Tür (33,0); Tisch Dienst (32,1); Transport Kraft (30,3); Tür Krieg (12,21); Preis Fehler (33,0); Berg Steuer (29,4); Rohr Zustand (28,5); Fisch Geist (22,10); Tier Staat (28,5); Fest Versicherung (31,2); Stadt Artikel (25,8); Zug Arzt (6,27); Tee Schule (33,0); Herz Unterricht (14,19); Zahn Lager (32,1); Atom Raum (32,1)

L1-R3: Left Position: Positive -ø- Bias; right constituent: Negative -ø- Bias:
Wand Sittich (30,3); Berg Hauptstadt (28,5); Wald Haushalt (26,7); Stadt Maßregel (25,8); Rohr Standard (25,8); Öl Geschenk (31,2); Zug Person (14,19); Fisch Hunger (28,5); Mond Kanzler (26,7); Transport Sekretär (24,9); Tür Klage (21,12); Stein Produktion (28,5); Zahn Kummer (32,1); Fest Koalition (30,3); Atom Moral (31,2); Herz Lotto (20,13); Tier Streit (28,5); Tee Bauch (33,0); Tisch Nest (27,6); Preis Verrat (31,2)

L2-R1: Left Position: Neutral -ø- Bias; Right Position: Positive -ø- Bias:
Meer Meter (5,28); Lamm Wächter (26,7); Kalb Analyse (4,29); Weide Bombe (5,28); Jahr Beere (0,33); Ohr Gelenk (17,16); Rebe Gerät (4,29); Kohle Säule (1,32); Mord Essen (8,25); Ei Kern (22,11); Fels Eisen (12,21); Alp Wolke (2,31); Watt Monat (26,7); Ausnahme Note (22,11); Achse Nummer (0,33); Zorn Schrift (12,21); Arzt Wolle (26,7); Verband Wurst (1,32); Tabak Flasche (23,10); Himmel Läufer (2,31)

L2-R2: Left Position: Neutral -∅- Bias; Right Position: Neutral -∅- Bias:
Kohle Krieg (5,28); Lamm Arzt (16,17); Jahr Boot (4,29); Ausnahme Unterricht (29,4); Ohr Steuer (13,20); Fels Dienst (7,26); Weide Zustand (13,20); Ei Artikel (19,14); Achse Raum (0,33); Arzt Kraft (27,6); Rebe Bett (1,32); Watt Geist (26,7); Kalb Lager (0,33); Mord Tür (10,23); Meer Sprache (3,30); Alp Schule (2,31); Zorn Spiegel (17,16); Tabak Staat (24,9); Verband Fehler (1,32); Himmel Versicherung (0,33)

L2-R3: Left Position: Neutral -∅- Bias; Right Position: Negative-∅- Bias:
Lamm Hunger (18,15); Weide Streit (8,25); Kohle Verrat (6,27); Verband Koalition (1,32); Ohr Kummer (6,27); Fels Haushalt (8,25); Ei Bauch (10,23); Tabak Nest (21,12); Zorn Moral (17,16); Mord Sekretär (13,20); Meer Hauptstadt (3,30); Watt Sittich (26,7); Jahr Geschenk (1,32); Ausnahme Kanzler (26,7); Arzt Klage (25,8); Alp Lotto (1,32); Kalb Maßregel (3,30); Achse Standard (0,33); Rebe Produktion (2,31); Himmel Person (0,33)

L3-R1: Left Position: Negative -∅- Bias; Right Position: Positive -∅- Bias:
Kanone Eisen (0,33); Maus Gelenk (18,15); Träne Beere (0,33); Zigarette Bombe (1,32); Suppe Analyse (1,32); Leiden Note (0,33); Geburt Wolle (1,32); Rose Flasche (0,33); Glocke Wolke (0,33); Hölle Kern (1,32); Treppe Säule (0,33); Mittag Wächter (0,33); Wolf Wurst (2,31); Rippe Nummer (0,33); Bauer Essen (2,31); Seife Gerät (0,33); Sonne Monat (0,33); Schiff Meter (4,29); Reich Läufer (3,30); Seemann Schrift (1,32)

L3-R2: Left Position: -∅- bias; Right Position: Neutral -∅- Bias:
Geburt Kraft (1,32); Bauer Unterricht (4,29); Wolf Krieg (1,32); Seife Boot (0,33); Maus Sprache (14,19); Schiff Zustand (2,31); Leiden Raum (0,33); Kanone Arzt (0,33); Sonne Dienst (0,33); Treppe Bett (0,33); Rose Steuer (0,33); Hölle Geist (0,33); Seemann Staat (0,33); Mittag Schule (0,33); Suppe Artikel (0,33); Glocke Tür (0,33); Reich Lager (4,29); Träne Spiegel (0,33); Zigarette Versicherung (0,33); Rippe Fehler (0,33)

L3-R3: Left Position: Negative -∅- Bias; Right Position: Negative -∅- Bias:
Glocke Produktion (0,33); Träne Sittich (0,33); Reich Hunger (1,32); Seemann Sekretär (2,31); Zigarette Haushalt (0,33); Kanone Lotto (0,33); Rose Nest (0,33); Treppe Streit (0,33); Suppe Kanzler (1,32); Geburt Geschenk (0,33); Seife Bauch

(0,33); Schiff Klage (1,32); Leiden Koalition (1,32); Mittag Kummer (0,33); Rippe Standard (0,33); Wolf Verrat (0,33); Sonne Hauptstadt (0,33); Bauer Maßregel (2,31); Maus Moral (16,17); Hölle Person (0,33)

Complex words in complex words

This chapter has been published as Andrea Krott, R. Harald Baayen, and Robert Schreuder· 1999, Complex words in complex words, *Linguistics* 37 (5), 905–926.

Abstract

Constituents of complex words can themselves be complex words. Some kinds of complex constituents appear more often than others. This study presents a quantitative investigation of this phenomenon. We show that many kinds of base words are significantly overrepresented or underrepresented. This holds not only for constituents of derived words, but also for constituents of compounds. We furthermore show that the degree of overrepresentation or underrepresentation correlates with word frequency, word length, and degree of productivity. We offer a functional explanation of this correlation in terms of processing and storage in the mental lexicon.

Introduction

It is well known that word formation rules accept several kinds of base words as input. As pointed out by Aronoff (1976), some kinds of base words of a given word formation rule give rise to more complex forms than others. He judged these differences in overall productivity important enough to warrant explicit mention in his formal definition of word formation rules. For the English prefix *un-*, e.g., he proposed the following rule in which the list of base words is "given roughly in order of productivity" (Aronoff 1976:63):

(20) *Rule of negative un#*

a. $[X]_{Adj} \rightarrow [un\#X]_{Adj}]_{Adj}$

semantics (roughly) $un\#X = not X$

b. Forms of the base

1. $X_V en$ (where *en* is the marker for past participle)
2. $X_V \#ing$
3. $X_V \#able$
4. $X + y$ (worthy)
5. $X + ly$ (seemly
6. $X\#ful$ (mindful)
7. $X - al$ (conditional)
8. $X\#like$ (warlike)

Corpus based data presented in Baayen & Renouf (1996) show that there are indeed substantial and significant differences in the numbers of base word types for *un-*. For instance, base words ending in *-ed* are very common, while base words ending in *-less* are virtually non-existent.

Given the fact that some kinds of base words occur more frequently than others, the following questions arise. First, are such unequal distributions simply reflections of the general proportions of complex words in the language? That many words in *-ed* and few words in *-less* give rise to *un-* formations would not be surprising at all if there would be many more independent words in *-ed* than in *-less* available in the language. We would only be dealing with a non-trivial phenomenon if there were relatively few formations in *-ed* and many formations in *-less*. In other words, further research is called for only if the distribution of base words for a particular

kind of complex word deviates significantly from the distribution of these words as independent words in the language

Second, if it is indeed the case that non-trivial unequal distributions exist in the domain of derivational morphology, the question arises whether similar unequal distributions can be observed in the domain of compounding as well

Third, if unequal distributions arise both in derivation and in compounding, then we are apparently dealing with a general phenomenon But why would this phenomenon exist? What kind of factors might give rise to such unequal distributions?

In what follows, we first examine the distribution of base words for the Dutch suffix *-heid*, a suffix similar to the English suffix *-ness* We introduce a statistical method for testing whether the distribution of base words differs from their distribution as independent words in the language We will show that indeed the two distributions differ significantly

We then extend our analysis to nominal compounds We again observe that the extent to which words from morphological categories are used as constituents in compounds differs remarkably from the extent to which these words are used on their own This suggests that we are indeed dealing with a general phenomenon

Finally, we will show that frequency of use, linguistic complexity, and degree of productivity are important factors underlying the observed patterns

Derived words in -heid

Table 7 1 in the Appendix summarizes various statistics for the different kinds of base words for the suffix *-heid* These statistics have been calculated using the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) This database contains frequency counts for some 120,000 morphologically analyzed lemmas based on a corpus of written Dutch of 42 million words The first part of the table lists the main derivational affixes that give rise to words in *-heid* The monomorphemic base words are labelled MONO, compounds are listed as COMP, and adjectivized participles are listed as PART The category listed as SEMI groups together those words in CELEX of doubtful morphological complexity (marked as I or U in CELEX, in what follows we will call this set semi-derived words) Finally, the remaining category SY contains almost exclusively synthetic compounds

The second column of Table 7 1 (labelled f) lists the number of types in *-heid* for these sets of base words The total number of formations in *-heid* is 2226, including 11 affixes not listed in Table 7 1 because they jointly account for 16 formations

only. Note that we have substantial variation. Base words in *-ig* (*groenig*, 'greenish') give rise to 255 *-heid* formations, while base words in *-s* (*schools*, 'schoolish') give rise to only 18 formations. The question that we now have to ask ourselves is whether these differences in the number of types are in any sense remarkable from a statistical point of view.

In the past, this question has been addressed by investigating either rival affixes (e.g., *-ness* and *-ity*, see Aronoff, 1976; Anshen & Aronoff, 1988) or a set of affixes sharing the same word category for the base word (Baayen & Renouf, 1996). The idea is that if a particular kind of base word gives rise to many formations in one affix and few formations in another affix, then, provided the difference is statistically reliable, we have genuine evidence that we are observing a non-trivial phenomenon worth further investigation.

In the present study we have opted for a different approach in which we compare for one kind of word formation the numbers of observed types for its various kinds of base words with the numbers that one would expect under chance conditions. To do so we make use of the binomial model. In the case of *-heid*, we regard the 2226 *-heid* formations as 2226 random trials. For a given kind of base word, we consider a trial to be successful if it yields a *-heid* formation with that particular kind of structure, i.e., if there is at least one token in our database for that particular type. (Note that the present statistical analysis has nothing to say about the token frequencies with which the individual types appear.) In other words, the f column in Table 7.1 can be viewed as listing the observed number of successes out of 2226 trials for each base word type.

How can we determine the expected number of successes? In the binomial scheme the expected number of successes equals np, where n denotes the number of trials and p the probability of success. In the case at hand, n is 2226. We can estimate p for a base type X by the relative type frequency of X in the list of all adjectives in CELEX which form the attested set of words to which *-heid* can be attached in principle.[1] There are 9925 such potential input words of which 528 belong to the morphological category of *-ig*. The column labelled *fcel* lists this number of types in CELEX for all base word types. We can now estimate the probability of success for *-ig* to be 528 / 9925 = 0.0532 and for *-s* to be 111 / 9925 = 0.0112. The

[1] Our counts of the number of types in CELEX to which *-heid* can attach in principle are raw counts. Our counts do not differentiate between base words for which a *-heid* formation is plausible versus implausible (Matthews, 1974:221–222), nor do they take possible semantic restrictions on the affixation of *-heid* (Bertram, Baayen, & Schreuder, 2000) into account. Here we simply assume that the effects of such constraints are uniformly distributed over the input domains. Further quantitative research is required here.

corresponding expected values are 0.0532 * 2226 = 118.42 and 0.0112 * 2226 = 24.90 respectively. Column E lists the expected numbers of types for all kinds of base words.

Comparing the observed and expected values, we observe far more *-ig* base words (255) than expected (118), while for *-s* the observed count (18) is smaller than expected (25). Are these differences between the observed and expected counts significant? Because the number of trials is large we can approximate the binomial model by a normal model and calculate Z-scores. To do so we need the standard deviation in addition to the expected counts. The standard deviation in the binomial model equals $\sqrt{np(1-p)}$, listed in Table 7.1 in column s. The Z-scores ($(f - np)/\sqrt{np(1-p)}$) are listed in column Z and the corresponding Bonferroni-adjusted significance levels in column $sign$ (*: .05; **: .01). Positive Z-scores imply overrepresentation, negative Z-scores imply underrepresentation. Table 7.1 shows that we have significant underrepresentation or overrepresentation for almost all base word types. The only exceptions are the adjectives in *-s* and the set of synthetic compounds. As a group, derived words are overrepresented as base words. The only affix that is significantly underrepresented is *-achtig*. The only other base word type exhibiting overrepresentation is the set of monomorphemic words. Significant underrepresentation is characteristic of compounds, participles, and semi-derived words.

We conclude that the phenomenon of overrepresentation and underrepresentation observed by Aronoff (1976), Anshen & Aronoff (1988), and Baayen & Renouf (1996) for English can also be observed for Dutch.

This phenomenon receives some qualitative support from the subset of *-heid* formations coined from adjectives in *-ig* (*groenigheid*, 'greenishness'). It has been observed that in some of these formations the suffix *-ig* does no longer contribute its own semantics: *stommig* means somewhat stupid, while *stommigheid* means 'stupidity'. This suggests that the sequence *-igheid* might be analyzed as a separate affix in its own right (Schultink, 1962; but see also De Haas & Trommelen, 1993, who do not make this distinction). If the combination of *-ig* and *-heid* is indeed developing into a single unit, then this provides qualitative evidence paralleling our quantitative evidence that the morphological structure of the base word in a complex word should be taken into account. Differences in over- and underrepresentation might then go hand in hand with subtle differences in semantics.

Compounds

Can we observe similar patterns of overrepresentation and underrepresentation for compounds? If we are dealing with a general phenomenon, one would expect that the left and right constituents of compounds behave in a similar way as the base words underlying formations in *-heid*. We have explored this possibility for Dutch and German nominal compounds using the CELEX lexical databases for Dutch and German. The German database lists some 52,000 entries based on a corpus of 6 million wordforms. Table 7.3 lists the same statistics as presented in Table 7.1 for a partition of left and right constituents into six kinds of base words: Monomorphemic base words (MONO), semi-derived words (SEMI), derived words (DER), compounds (COMP), synthetic compounds (SY), and a small heterogeneous set of other kinds of complex words (O). In both languages none of these kinds of base words occur with frequencies that one would expect under chance conditions, as shown by the Z-scores and the associated probabilities. Just as for *-heid*, monomorphemic words are strongly overrepresented, while the compounds and to a lesser degree the synthetic compounds are underrepresented. Dutch and German diverge with respect to the set of derived words. In Dutch, derived words are overrepresented, while in German they are underrepresented. Interestingly, left and right constituents reveal exactly the same pattern, even though the right headedness of most compounds might have led to an asymmetry.

The role of word frequency

Is there any systematicity in the patterns of overrepresentation and underrepresentation observed in the previous section? Altmann (1988) suggests that higher frequency words are more likely to appear as constituents in compounds than lower frequency words. If this hypothesis generalizes to complex words in general, the following relation might hold:

> The higher the average word frequency for a given base word type, the higher the chance of it being overrepresented in complex words.

To test this hypothesis, we calculated the mean log frequency using the CELEX lexical database for each base word type, the column labelled $meanf$ in Tables 7.1–7.3.[2] Figures 7.1–7.2 show that we indeed have a positive correlation between

[2]The logarithmic transformation largely eliminates the Zipfian skewness from the word frequency distributions and allows us to gauge more precisely the central tendency in the data. In addition, the

mean log frequency and Z-score. Figure 7.1 presents a scatterplot for the *-heid* data. Monomorphemic words have the highest mean log frequency and the highest positive Z-score, while compounds have a low mean log frequency and a large negative Z-score. The other kinds of base words are scattered between these extremes. Both a Pearson correlation analysis and a Spearman rank correlation analysis show that the correlation between mean log frequency and Z-score is reliable (r = .58, t(13) = 2.56, p = .024; r_S = .53, p = .049). The solid line in Figure 7.1 represents the corresponding mean squares regression line.

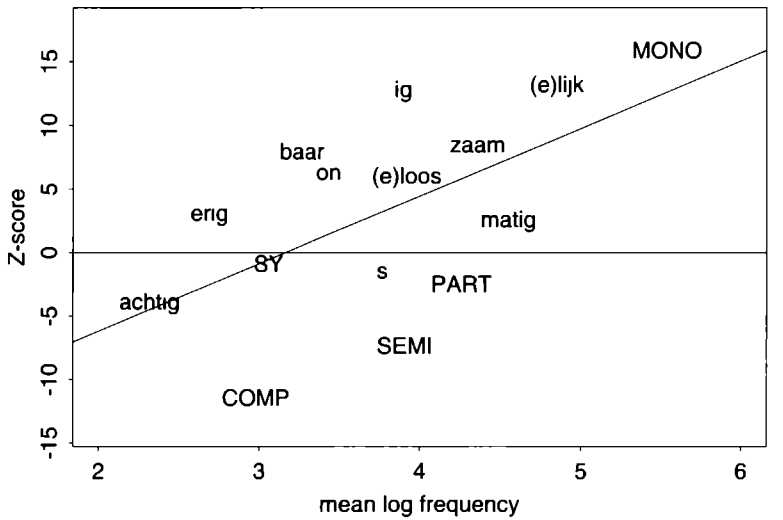


Figure 7.1: Mean log frequency and Z-score for base word types of *-heid* formations with mean squares regression line.

Figure 7.2 presents similar scatterplots for the left and right constituents of Dutch and German compounds. As before, the monomorphemic words appear in the upper right corners of the scatterplots and the compounds in the lower left corners. Despite the small number of base word categories, the correlations between mean log frequency and Z-score are all reliable (left constituents Dutch: r = .91, t(4) = 4.41, p = .012; r_S = 1, p = .030; right constituents Dutch: r = .91, t(4) = 4.41, p = .012; r_S = 1, p = .030; left constituents German: r = .92, t(4) = 4.55, p = .011;

human processing system is also sensitive to log frequency rather than absolute frequency.

r_S = .94, p = .041; right constituents German: r = .89, t(4) = 3.96, p = .017; r_S = .94, p = .041).

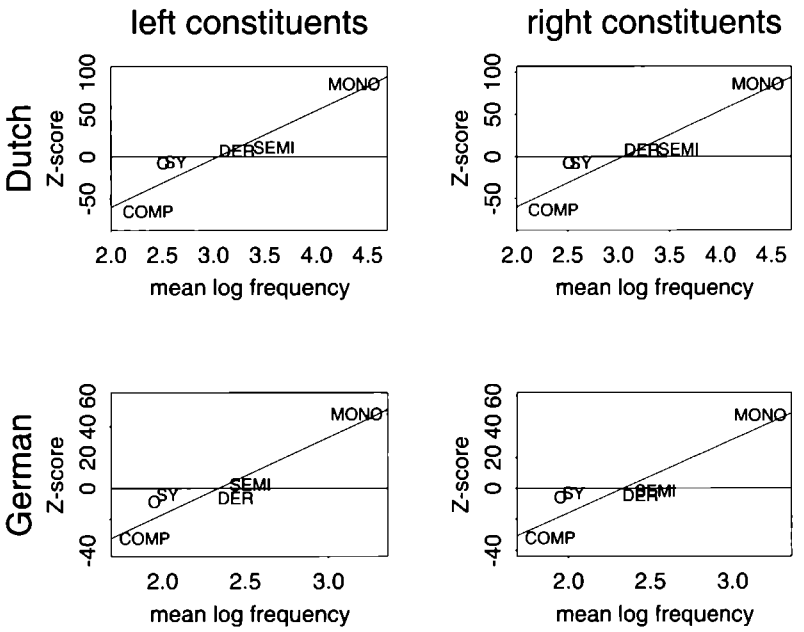


Figure 7.2: Mean log frequency and Z-score for base word types of Dutch and German compounds with mean squares regression lines.

Thus far, the data support our hypothesis that word frequency is an important factor co-determining the extent to which base words appear in complex words. As a final test, we calculated the mean log frequency and the Z-scores for the various kinds of derived words that appear as left and right constituents in Dutch compounds. Tables 7.2–7.4 and Figure 7.3 summerize the results. The scatterplots reveal some outliers, notably the nominalizing suffixes *-ing* ('-ing') and *-atie* ('-ation') in the upper panel, and the nominalizing suffixe *-ing* ('-ing'), *-er* ('-er'), and *-heid* ('-ness') in the lower panel. Given this outlier structure, we have only calculated the Spearman rank correlations, which again show that we are dealing with reliable correlations (left derivations: r_S = .71, p < .0001; right derivations: r_S = .47, p = .007). The solid lines in Figure 7.3 represent the least median squares regression lines.

The role of word length

We have shown that the average word frequency of a particular kind of base word is an important factor co-determining its use in complex words. It is well known that word frequency is strongly correlated with word length. To show that this relation also holds for constituents in complex words, we divided the Dutch data in classes of different lengths. Table 7.5a lists the classes for left compound constituents when measuring length in terms of number of morphemes. Table 7.5b lists the classes when measuring length in terms of number of phonemes. In both tables, the column labelled f contains the number of words in each class, and the column $mean f$ lists their mean log frequency. Comparable data for base words used in *-heid* formations and for right constituents of compounds are presented in Table 7.6 and Table 7.7.

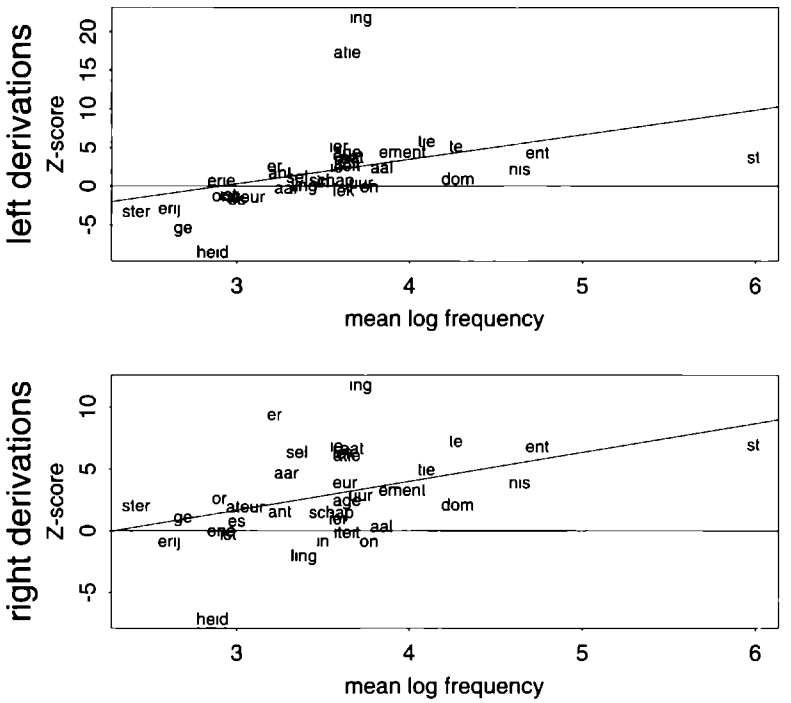


Figure 7.3: Mean log frequency and Z-score for left and right derivational constituent types of Dutch compounds with median squares regression lines.

To illustrate the strong negative correlation between length and frequency, we consider the left constituents of compounds in some more detail. The top left panel

of Figure 7.4 plots mean log frequency as a function of number of phonemes (r_s = -.99; p $<$ .0001).

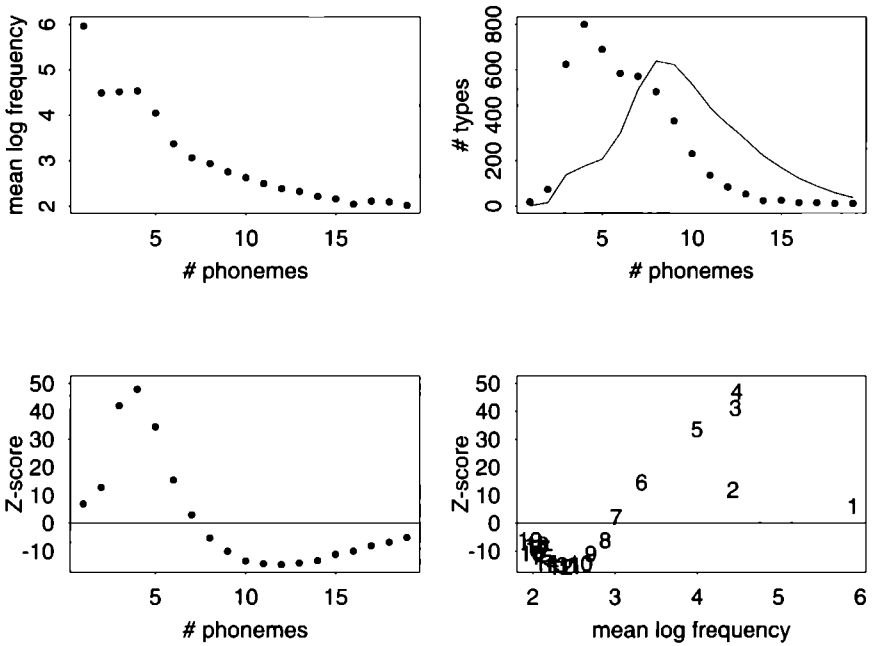


Figure 7.4: The relation between mean constituent length, number of constituent types, mean frequency of occurrence, and Z-score for left constituents of Dutch compounds. The dots in the upper right panel represent the observed numbers of constituent types, the solid line represents the corresponding expected number of types. The numbers in the lower right panel represent word length in phonemes.

Given this negative correlation between word frequency and word length, we also expect the following relation to hold:

> The longer a base word, the higher the chance of it being underrepresented in complex words.

To test this hypothesis, we calculated for each length class a Z-score, as we did for the constituent types in the previous section. The results of the Z-score statistics are listed in Tables 7.5–7.7. As expected, the Z-scores reveal that both short base words and short compound constituents are indeed overrepresented, while long base words and long compound constituents are underrepresented. The top

right panel of Figure 7.4 shows for the left constituents of compounds how the number of types in each phonemic length class (represented by dots) diverge from the expected number of types (represented by a solid line). For word length 1 the observed and expected number of types are nearly identical. For word lengths 2–7 the observed number of types exceeds the expected number of types, especially for word lengths 3–6. From lengths 8–19 the observed number of types is smaller than the expected number of types, especially for lengths 9–15. Note that there are relatively few types with very small or very large word length. We see the same pattern in the lower left panel of Figure 7.4 which plots the corresponding Z-scores as a function of word length.

The lower right panel of Figure 7.4 plots the length classes in the plane spanned by mean log frequency and Z-score. The underrepresented sets of constituents consist of words which are infrequent and long, while the overrepresented sets of constituents consist of words which are frequent and short. In sum, constituents in complex words reveal a correlational system in which word length, mean log frequency, and number of types are all interrelated.[3]

A productivity paradox

We have seen that word frequency and word length co-determine how often complex words appear as constituents in other complex words. Especially short and frequent words give rise to overrepresentation. Paradoxically, this suggests that those categories of base words that have a low category-conditioned degree of productivity are relatively more productive as constituents in other complex words than base words that have a high category-conditioned degree of productivity. The category-conditioned degree of productivity is defined as follows (Baayen 1992; see Baayen, 1994, for experimental evidence):

$$\mathcal{P} = \frac{V(1,N)}{N}, \tag{7.1}$$

with $V(1,N)$ the number of hapax legomena (types occurring once only) in a sample of N tokens of a given category. This statistic estimates the probability of sampling a word that has not yet been observed in the previous N tokens of the morphological category. Thus, a base word category with 1000 tokens and

[3] In the presented data, word frequency and length are so strongly correlated that it proved to be impossible to ascertain the extent to which these factors might play an independent role.

50 hapax legomena has a category-conditioned degree of productivity equal to $\mathcal{P}$ = .05. Another category with 10000 tokens and 50 hapax legomena has a category-conditioned degree of productivity equal to $\mathcal{P}$ = .005. Note that the probability of sampling new unobserved types decreases as N increases. A category with many short and high-frequency words will have a large value of N and hence a lower $\mathcal{P}$ compared to a category with only a few high-frequency forms. This leads to the following paradox:

> The more productive an affix, the greater the degree to which it is underrepresented in other complex words. The less productive an affix, the more it is overrepresented in other complex words.

In other words, the relative productivity of an affix, i.e., the degree to which it is overrepresented, is negatively correlated with its category-conditioned degree of productivity.

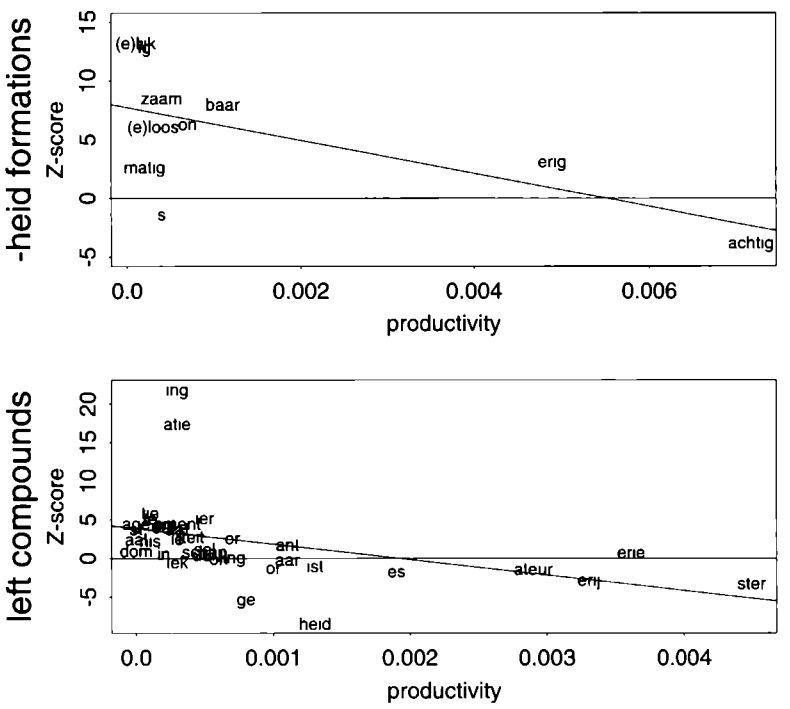


Figure 7.5: Degree of productivity and Z-score for base word types of *-heid* formation and left constituents of Dutch compounds.

To test this prediction, we first investigated the relation between underrepresentation and overrepresentation expressed in Z-scores with estimates of the category-conditioned degree of productivity.[4] Figure 7.5 plots categories in the plane of $\mathcal{P}$ and Z for base words of *-heid* formations (upper panel) and for left constituents of Dutch compounds (lower panel). The particular values of $\mathcal{P}$ are listed in Table 7.1 and Table 7.2 in the column labelled $prod$.

For the base categories of words in *-heid* we observe a trend in the expected direction. The category with the highest $\mathcal{P}$-value (*-achtig*, '-like') has the lowest Z-score. Conversely, the category with the lowest $\mathcal{P}$-value (*-(e)lijk*, '-able') has the highest Z-score. However, due to the small number of observations, the Spearman rank correlation is not fully reliable (r_s = -.52; p = .06, one-tailed test). Interestingly, the object-modifying rival affixes *-(e)lijk* (*verwerpelijk*, 'objectionable') and *-baar* (*toepasbaar*, 'applicable') behave exactly as expected. Van Marle (1988) and Hüning & Van Santen (1994) point out that *-baar* is productive and semantically transparent, while *-(e)lijk* is unproductive and appears in many semantically opaque words. This difference is reflected in the $\mathcal{P}$-values of these suffixes, and indeed we observe that *-(e)lijk* has the higher Z-score.

For the base categories appearing as left constituents in compounds (shown in the lower panel of Figure 7.5) we observe a very clear negative correlation between category-conditioned degree of productivity and Z-score (r_s = -.69, p = .0001): the more productive categories have the lower Z-scores. These data show that word frequency and word length have to be considered in combination with degree of productivity when studying the contribution of morphological categories to the productivity of other complex words.

General Discussion

The aim of this paper has been to study the extent to which the productivity of derivation and compounding is influenced by the morphological structure of base words. We have first shown that the unequal contributions of different kinds of base words are extremely unlikely to be a chance phenomenon. We have further shown that the phenomenon of unequal contributions is not limited to derivation, but that it likewise occurs in the domain of compounding, both for left and right constituents.

[4]The CELEX lexical database does not provide counts of hapax legomena. We have therefore approximated the category-conditioned degree of productivity by the ratio of dis legomena (words occurring twice) to the total number of tokens of a category in CELEX.

Finally, we have shown that the extent to which particular kinds of base words are overrepresented or underrepresented correlates with their mean frequency of use and their length (measured in number of phonemes or morphemes). Shorter and more frequent words are overrepresented, longer and less frequent words are underrepresented. Paradoxically, categories with a low degree of productivity are relatively more productive as constituents in other complex words.

The correlation of word frequency, word length, and category-conditioned degree of productivity on the one hand with the degree of overrepresentation (Z-scores) on the other hand explains 1/5 up to 1/3 of the variance in the data. This observation raises the following question. How can we understand this non-trivial role of word frequency, word length, and productivity as explanatory variables?

In all our calculations of expected numbers of types, we have assumed the null-hypothesis that all word types are equiprobable. The observed underrepresentation and overrepresentation show that this null-hypothesis is incorrect. This raises the question in what way some words are more likely to be selected as a constituent than other words. From a psycholinguistic point of view, we can understand the finding that base word categories which comprise frequent words are overrepresented compared to categories comprising less frequent words in terms of the word frequency effect (e.g., Scarborough, Cortese, & Scarborough, 1977; Hasher & Zacks, 1984). The word frequency effect is the finding that higher frequency words are recognized and produced more quickly and accurately than lower frequency words. Assuming that a wide range of complex words is stored in the mental lexicon, the same word frequency effect applies to complex words as well (Baayen, Dijkstra, & Schreuder, 1997; Sereno & Jongman, 1997). This means that higher frequency complex words are more accessible as potential constituents than lower frequency words. A category of base words that contains many frequent formations will then be overrepresented.

Similarly, shorter words are easier to produce and recognize than longer words (e.g., Henderson, 1985, p. 470-471). Since higher frequency words tend to have more meanings and shades of meanings (Köhler, 1986; Altmann, Beöthy, & Best, 1982; Paivio, Yuille, & Madigan, 1968; Reder, Anderson, & Bjork, 1974), they are also more likely to be selected during the process of conceptualization and lexical selection in speech production. Note, furthermore, that less productive and unproductive categories typically comprise higher frequency formations that tend to have more, and more opaque meanings. Such formations have to be stored in the mental lexicon in any case where they are readily available for further word formation. This

explains the paradox that less productive categories are relatively more productive as constituents, a paradox that is entirely unexpected on the basis of the combinatorial properties of word formation rules only. From this perspective, any summary description of a word formation rule is incomplete without a quantitative description of the pattern of overrepresentation and underrepresentation of its base words.

In traditional analyses of morphological productivity, the role of phonological, semantic and syntactic constraints has figured prominently (Van Marle, 1985; Booij, 1977). The morphological restrictions formalized by Aronoff (1976) as part of generative word formation rules have received little attention. The present results, however, show that these morphological restrictions are statistically non-trivial: constituent length, constituent frequency, and the productivity of the morphological category to which the constituent belongs form a correlational complex that codetermines the overall productivity of a word formation rule. We have offered a quantitative, partial explanation in terms of the mental lexicon, but further qualitative research is neccessary in order to fully understand how such morphological restrictions arise.

References

Altmann, G.: 1988, Hypotheses about compounds, *in* R. Hammerl (ed.), *Glottometrika 10*, Brockmeyer, Bochum, pp. 100–107.

Altmann, G., Beöthy, E. and Best, K.-H.: 1982, Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz (Meaning complexity and Menzerath's law), *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* **35**(5), 537–543.

Anshen, F. and Aronoff, M.: 1988, Producing morphologically complex words, *Linguistics* **26**, 641–655.

Aronoff, M.: 1976, *Word Formation in Generative Grammar*, MIT Press, Cambridge, Mass.

Baayen, R. H.: 1992, Quantitative aspects of morphological productivity, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1991*, Kluwer Academic Publishers, Dordrecht, pp. 109–149.

Baayen, R. H.: 1994, Productivity in language production, *Language and Cognitive Processes* **9**, 447–469.

Baayen, R. H. and Renouf, A.: 1996, Chronicling The Times: productive lexical innovations in an English newspaper, *Language* **72**, 69–96.

Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **36**, 94–117.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bertram, R., Baayen, R. H. and Schreuder, R.: 2000, Effects of family size for complex words, *Journal of Memory and Language* **42**, 390–405.

Booij, G. E.: 1977, *Dutch Morphology. A Study of Word Formation in Generative Grammar*, Foris, Dordrecht.

De Haas, W. and Trommelen, M.: 1993, *Morfologisch handboek van het Nederlands* (Morphological handbook of Dutch), SDU, Den Haag.

Hasher, L. and Zacks, R. T.: 1984, Automatic processing of fundamental information. The case of frequency of occurrence, *American Psychologist* **39**, 1372–1388.

Henderson, L.: 1985, Issues in the modelling of pronunciation assembly in normal

reading, *in* K. Patterson, J. Marshall and M. Coltheart (eds), *Surface Dyslexia: Neuropsychological and Cognitive Studies on Phonological Reading*, Lawrence Erlbaum, London, pp. 459–508.

Hüning, M. and Van Santen, A.. 1994, Productiviteitsveranderingen: de adjectieven op *-lijk* en *-baar* (Changes of productivity: adjectives ending in *-lijk* and *-baar*), *Leuvense Bijdragen* **83**(1), 1–29.

Köhler, R.: 1986, *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*, Brockmeyer, Bochum.

Marle, J. v.: 1985, *On the Paradigmatic Dimension of Morphological Creativity*, Foris, Dordrecht.

Marle, J. v.: 1988, Betekenis als factor bij produktiviteitsverandering (iets over de deverbale categorieën op *-lijk* en *-baar*) (Meaning as factor for change in productivity. About the deverbal categories ending in *-lijk* and *-baar*), *Spektator* **17**, 341–359.

Matthews, P.: 1974, *Morphology. An Introduction to the Theory of Word Structure*, Cambridge University Press, London.

Paivio, A., Yuille, J. C. and Madigan, S.: 1968, Concreteness, imagery, and meaningness values for 925 nouns, *Journal of Experimental Psychology Monograph* **76** (I), Pt.2.

Reder, L. M., Anderson, J. R. and Bjork, R. A.: 1974, A semantic interpretation of encoding specificity, *Journal of Experimental Psychology* **102**, 648–656.

Scarborough, D. L., Cortese, C. and Scarborough, H. S.: 1977, Frequency and repetition effects in lexical memory, *Journal of Experimental Psychology: Human Perception and Performance* **3**, 1–17.

Schultink, H.: 1962, *De Morfologische Valentie van het Ongelede Adjectief in Modern Nederlands (Morphological valency of the simplex adjective in modern Dutch)*, Van Goor & Zonen, Den Haag.

Sereno, J. and Jongman, A.: 1997, Processing of English inflectional morphology, *Memory and Cognition* **25**, 425–437.

Appendix

Table 7.1: **Base word classes of -heid formations**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); prod: category-conditioned degree of productivity; SEMI: doubtful morphologically complex words; MONO: monomorphemic words; SY: synthetic compounds; COMP: compounds consisting of two nouns and a possible linking morpheme; PART: present participle.

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
on-	264	3.44	812	0.0818	182.1	12.9	6.3	**	0.0007
-ig	255	3.90	528	0.0532	118.4	10.6	12.9	**	0.0002
-(e)lijk	188	4.85	335	0.0338	75.1	8.5	13.3	**	0.0001
-baar	91	3.27	181	0.0182	40.6	6.3	8.0	**	0.0011
-(e)loos	71	3.92	157	0.0158	35.2	5.9	6.1	**	0.0003
-erig	46	2.69	130	0.0131	29.2	5.4	3.1	*	0.0049
-zaam	30	4.36	32	0.0032	7.2	2.7	8.5	**	0.0004
-achtig	28	2.32	251	0.0253	56.3	7.4	-3.8	**	0.0072
-s	18	3.77	111	0.0112	24.9	5.0	-1.4		0.0004
-matig	9	4.55	17	0.0017	3.8	2.0	2.7		0.0002
SEMI	313	3.91	2015	0.2030	451.9	19.0	-7.3	**	0.0002
MONO	311	5.54	593	0.0597	133.0	11.2	15.9	**	0.0000
SY	249	3.06	1169	0.1178	262.2	15.2	-0.9		0.0018
PART	246	4.26	1270	0.1280	284.8	15.8	-2.5		0.0001
COMP	91	2.98	1184	0.1193	265.6	15.3	-11.4	**	0.0012

$n = \sum f = 2226$

Table 7.2: **Derivation classes in Dutch left compound constituents**: $f1$: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); prod: category-conditioned degree of productivity.

a. Dutch left compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	2592	4.37	5180	0.0952	592.2	23.2	86.4	**
SEMI	1567	3.59	10647	0.1957	1217.3	31.3	11.2	**
DER	1335	3.23	9911	0.1822	1133.1	30.4	6.6	**
COMP	658	2.36	28168	0.5178	3220.5	39.4	-65.0	**
SY	45	2.63	970	0.0178	110.9	10.4	-6.3	**
O	23	2.51	876	0.0161	100.2	9.9	-7.8	**

$n = \sum f = 6220$

b. Dutch right compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	2342	4.37	5180	0.0952	502.4	21.3	86.3	**
SEMI	1288	3.59	10647	0.1957	1032.6	28.8	8.9	**
DER	1184	3.23	9911	0.1822	961.2	28.0	8.0	**
COMP	428	2.36	28168	0.5178	2731.7	36.3	-63.5	**
SY	20	2.63	970	0.0178	94.1	9.6	-7.7	**
O	14	2.51	876	0.0161	85.0	9.1	-7.8	**

$n = \sum f = 5276$

c. German left compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	1534	3.17	4355	0.2101	539.0	20.6	48.2	**
DER	579	2.44	5849	0.2822	724.0	22.8	-6.4	**
SEMI	283	2.53	1964	0.0947	243.1	14.8	2.7	*
COMP	155	1.89	7638	0.3685	945.5	24.4	-32.4	**
O	14	1.95	779	0.0376	96.4	9.6	-8.6	**
SY	1	2.03	145	0.0070	18.0	4.2	-4.0	**

$n = \sum f = 2566$

d. German right compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	1401	3.17	4355	0.2101	478.8	19.5	47.4	**
DER	546	2.44	5849	0.2822	643.0	21.5	-4.5	**
SEMI	191	2.53	1964	0.0947	215.9	14.0	-1.8	
COMP	104	1.89	7638	0.3685	839.7	23.0	-32.0	**
O	33	1.95	779	0.0376	85.6	9.1	-5.8	**
SY	4	2.03	145	0.0070	15.9	4.0	-3.0	**

$n = \sum f = 2279$

Table 7.3: **Compound constituents**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); MONO: monomorphemic words; SEMI: doubtful morphologically complex words; DER: derived words; COMP: compounds consisting of two nouns and a possible linking morpheme; SY: synthetic compounds; O: remaining words not belonging to any of the other classes.

class	$f1$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-ing	551	3.72	1986	0.0365	227.1	14.8	21.9	**	0.0003
-atie	156	3.64	373	0.0069	42.7	6.5	17.4	**	0.0003
-er	127	3.22	881	0.0162	100.7	10.0	2.6		0.0007
-heid	83	2.86	1759	0.0323	201.1	14.0	-8.5	**	0.0013
-ie	48	3.58	292	0.0054	33.4	5.8	2.5		0.0003
-iteit	30	3.64	159	0.0029	18.2	4.3	2.8		0.0004
-te	22	4.27	67	0.0012	7.7	2.8	5.2	**	0.0001
-tie	20	4.10	51	0.0009	5.8	2.4	5.9	**	0.0001
-sel	18	3.35	118	0.0022	13.5	3.7	1.2		0.0005
-schap	16	3.55	113	0.0021	12.9	3.6	0.9		0.0005
-ier	16	3.59	41	0.0008	4.7	2.2	5.2	**	0.0005
-aat	16	3.67	56	0.0010	6.4	2.5	3.8	**	0.0003
-eur	16	3.63	53	0.0010	6.1	2.5	4.0	**	0.0002
ge-	15	2.69	468	0.0086	53.5	7.3	-5.3	**	0.0008
-aar	14	3.29	128	0.0024	14.6	3.8	-0.2		0.0011
-age	13	3.64	35	0.0006	4.0	2.0	4.5	**	0.0000
-ant	12	3.25	63	0.0012	7.2	2.7	1.8		0.0011
-ling	11	3.39	92	0.0017	10.5	3.2	0.2		0.0007
-ent	11	4.74	28	0.0005	3.2	1.8	4.4	**	0.0002
-ement	11	3.96	27	0.0005	3.1	1.8	4.5	**	0.0003
-st	9	5.99	24	0.0004	2.7	1.7	3.8	**	0.0000
-iek	8	3.62	83	0.0015	9.5	3.1	-0.5		0.0003

Table 7.3 (continued)

class	$f1$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-nis	8	4.64	32	0.0006	3.7	1.9	2.3		0.0001
-erij	7	2.61	169	0.0031	19.3	4.4	-2.8		0.0033
on-	7	3.77	62	0.0011	7.1	2.7	-0.1		0.0006
-ist	7	2.95	90	0.0017	10.3	3.2	-1.0		0.0013
-in	7	3.50	49	0.0009	5.6	2.4	0.6		0.0002
-ster	3	2.42	139	0.0026	15.9	4.0	-3.2	*	0.0045
-or	3	2.90	52	0.0010	6.0	2.4	-1.2		0.0010
-uur	3	3.72	20	0.0004	2.3	1.5	0.5		0.0005
-dom	3	4.28	15	0.0003	1.7	1.3	1.0		0.0000
-aal	3	3.84	7	0.0001	0.8	0.9	2.5		0.0000
-es	2	3.00	53	0.0010	6.1	2.5	-1.7		0.0019
-erie	2	2.91	10	0.0002	1.1	1.1	0.8		0.0036
-ateur	0	3.05	16	0.0003	1.8	1.4	-1.4		0.0029

$n = \sum f1 = 1278$

Table 7.4: **Derivation classes in Dutch right compound constituents**: $f2$: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); prod: category-conditioned degree of productivity.

class	$f2$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-ing	354	3.72	1986	0.0365	192.6	13.6	11.9	**	0.0003
-er	172	3.22	881	0.0162	85.4	9.2	9.4	**	0.0007
-heid	79	2.86	1759	0.0323	170.6	12.9	-7.1	**	0.0013
-atie	73	3.64	373	0.0069	36.2	6.0	6.1	**	0.0003
-ie	65	3.58	292	0.0054	28.3	5.3	6.9	**	0.0003
ge-	53	2.69	468	0.0086	45.4	6.7	1.1		0.0008
-sel	33	3.35	118	0.0022	11.4	3.4	6.4	**	0.0005
-aar	29	3.29	128	0.0024	12.4	3.5	4.7	**	0.0011
-iek	26	3.62	83	0.0015	8.1	2.8	6.3	**	0.0003
-te	25	4.27	67	0.0012	6.5	2.6	7.3	**	0.0001
-ster	21	2.42	139	0.0026	13.4	3.7	2.1		0.0045
-aat	21	3.67	56	0.0010	5.4	2.3	6.7	**	0.0003
-schap	16	3.55	113	0.0021	11.0	3.3	1.5		0.0005
-tie	16	4.10	51	0.0009	5.0	2.2	5.0	**	0.0001
-iteit	15	3.64	159	0.0029	15.4	3.9	-0.1		0.0004
-ent	14	4.74	28	0.0005	2.7	1.7	6.9	**	0.0002
-eur	14	3.63	53	0.0010	5.1	2.3	3.9	**	0.0002
-erij	13	2.61	169	0.0031	16.4	4.0	-0.8		0.0033
-st	13	5.99	24	0.0004	2.3	1.5	7.0	**	0.0000

Table 7.4 (continued)

class	$f2$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-or	11	2.90	52	0.0010	5.0	2.2	2.7		0.0010
-ant	10	3.25	63	0.0012	6.1	2.5	1.6		0.0011
-nis	10	4.64	32	0.0006	3.1	1.8	3.9	**	0.0001
-ist	8	2.95	90	0.0017	8.7	3.0	-0.3		0.0013
-ement	8	3.96	27	0.0005	2.6	1.6	3.3	*	0.0003
-age	8	3.64	35	0.0006	3.4	1.8	2.5		0.0000
-es	7	3.00	53	0.0010	5.1	2.3	0.8		0.0019
-ier	6	3.59	41	0.0008	4.0	2.0	1.0		0.0005
-uur	6	3.72	20	0.0004	1.9	1.4	2.9		0.0005
on-	4	3.77	62	0.0011	6.0	2.5	-0.8		0.0006
-ateur	4	3.05	16	0.0003	1.6	1.3	2.0		0.0029
-dom	4	4.28	15	0.0003	1.5	1.2	2.1		0.0000
-ling	3	3.39	92	0.0017	8.9	3.0	-2.0		0.0007
-in	3	3.50	49	0.0009	4.8	2.2	-0.8		0.0002
-erie	1	2.91	10	0.0002	1.0	1.0	0.0		0.0036
-aal	1	3.84	7	0.0001	0.7	0.8	0.4		0.0000

$n = \sum f2 = 1146$

Table 7.5: **Length of left Dutch compound constituents**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of types in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01).

a. Morphemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	2591	4.37	5192	0.0954	593.6	23.2	86.2	**
2	1936	2.69	28674	0.5271	3278.4	39.4	-34.1	**
3	117	2.24	9474	0.1741	1083.2	29.9	-32.3	**
4	5	2.48	408	0.0075	46.7	6.8	-6.1	**

$n = \sum f = 4649$

b. Phonemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	8	5.91	11	0.0002	1.3	1.1	6.0	**
2	63	4.44	138	0.0025	15.8	4.0	11.9	**
3	615	4.47	1204	0.0221	137.7	11.6	41.1	**
4	790	4.49	1539	0.0283	176.0	13.1	47.0	**
5	680	4.00	1804	0.0332	206.3	14.1	33.6	**
6	575	3.32	2809	0.0516	321.2	17.5	14.5	**
7	561	3.01	4491	0.0826	513.5	21.7	2.2	
8	493	2.88	5579	0.1025	637.9	23.9	-6.1	**
9	365	2.70	5452	0.1002	623.3	23.7	-10.9	**
10	220	2.57	4697	0.0863	537.0	22.2	-14.3	**
11	125	2.45	3798	0.0698	434.2	20.1	-15.4	**
12	73	2.33	3144	0.0578	359.5	18.4	-15.6	**
13	42	2.28	2563	0.0471	293.0	16.7	-15.0	**
14	14	2.17	1946	0.0358	222.5	14.7	-14.2	**
15	15	2.11	1465	0.0269	167.5	12.8	-11.9	**
16	4	1.99	1073	0.0197	122.7	11.0	-10.8	**
17	4	2.06	760	0.0140	86.9	9.3	-9.0	**
18	1	2.04	516	0.0096	59.0	7.6	-7.6	**
19	1	1.96	330	0.0061	37.7	6.1	-6.0	**

$n = \sum f = 4649$

Table 7.6: **Length of Dutch right compound constituents**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of types in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01).

a. Morphemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	2343	4.37	5192	0.0954	503.5	21.3	86.2	**
2	1564	2.69	28674	0.5271	2780.8	36.3	-33.6	**
3	74	2.24	9474	0.1741	918.8	27.6	-30.7	**
4	3	2.48	408	0.0075	39.6	6.3	-5.8	**

$n = \sum f = 3984$

b. Phonemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	3	5.91	11	0.0002	1.1	1.0	1.9	
2	46	4.44	138	0.0025	13.4	3.7	8.9	**
3	588	4.47	1204	0.0221	116.8	10.7	44.1	**
4	775	4.49	1539	0.0283	149.3	12.0	52.0	**
5	659	4.00	1804	0.0332	175.0	13.0	37.2	**
6	500	3.32	2809	0.0516	272.4	16.1	14.2	**
7	409	3.01	4491	0.0826	435.5	20.0	-1.3	
8	366	2.88	5579	0.1025	541.1	22.0	-7.9	**
9	279	2.70	5452	0.1002	528.7	21.8	-11.5	**
10	173	2.57	4697	0.0863	455.5	20.4	-13.9	**
11	84	2.45	3798	0.0698	368.3	18.5	-15.4	**
12	44	2.33	3144	0.0578	304.9	17.0	-15.4	**
13	22	2.28	2563	0.0471	248.6	15.4	-14.7	**
14	12	2.17	1946	0.0358	188.7	13.5	-13.1	**
15	11	2.11	1465	0.0269	142.1	11.8	-11.2	**
16	5	1.99	1073	0.0197	104.1	10.1	-9.8	**
17	4	2.06	760	0.0140	73.7	8.5	-8.2	**
18	1	2.04	516	0.0095	50.0	7.0	-7.0	**
19	2	1.96	330	0.0061	32.0	5.6	-5.3	**
21	1	1.90	133	0.0024	12.9	3.6	-3.3	**

$n = \sum f = 3984$

Table 7 7 **Length of base words of *-heid* formations** f number of types, $meanf$ mean log token frequency, $fcel$ number of types in CELEX, p probability of a word being a member of the class, E expected number of types, s standard deviation, Z Z-score, $sign$ Bonferroni-adjusted significance level (* 05, ** 01)

a. Morphemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	311	5 54	593	0 0597	133 0	11 2	15 9	**
2	1107	3 44	4921	0 4958	1103 7	23 6	0 1	
3	247	3 05	1109	0 1117	248 7	14 9	-0 1	
4	2	2 55	17	0 0017	3 8	2 0	-0 9	

$n = \sum f = 1667$

b. Phonemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
2	6	6 25	21	0 0021	4 7	2 2	0 6	
3	112	5 86	186	0 0187	41 7	6 4	11 0	**
4	118	5 38	218	0 0220	48 9	6 9	10 0	**
5	141	4 47	323	0 0325	72 4	8 4	8 2	**
6	140	3 81	498	0 0502	111 7	10 3	2 8	*
7	189	3 71	720	0 0725	161 5	12 2	2 3	
8	265	3 51	1049	0 1057	235 3	14 5	2 1	
9	243	3 38	1040	0 1048	233 2	14 5	0 7	
10	211	3 30	928	0 0935	208 1	13 7	0 2	
11	125	3 11	673	0 0678	150 9	11 9	-2 2	
12	74	2 30	442	0 0445	99 1	9 7	-2 6	
13	22	3 06	227	0 0229	50 9	7 1	-4 1	**
14	11	2 83	145	0 0146	32 5	5 7	-3 8	**
15	6	2 80	83	0 0084	18 6	4 3	-2 9	*
16	4	2 38	40	0 0040	9 0	3 0	-1 7	

$n = \sum f = 1667$

The function of Dutch linking elements

This chapter will be published as Andrea Krott, Robert Schreuder, and R. Harald Baayen: A note on the function of Dutch linking elements, *Yearbook of Morphology*.

Abstract

This study addresses the question of the functionality of linking elements in Dutch noun-noun compounds when they follow derived left constituents. In particular, we focus on the possible function of opening derived words ending in closing suffixes for further word formation (Aronoff & Fuhrhop, submitted). We address this question by means of a statistical analysis of the distributional properties of compounds and their constituents. We present evidence that both the linking *-s-* and the linking *-en-* open suffixes for further word formation. Prototypical closing suffixes, however, are only opened by the linking element *-s-*. In addition, we show that this functional link between suffixes and linking elements breaks up general distributional properties of derived forms that occur as left compound constituents.

Introduction

Dutch noun-noun compounds often contain linking elements, namely *-s-* (e.g., *schaap+s+kooi* 'sheep fold') and *-en-* (e.g., *boek+en+kast* 'book shelf') or its orthographic variant *-e-* (*zonn+e+schijn* 'sun shine'). Of the 23,000 Dutch noun-noun compounds that are listed in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), 31% contain either a linking *-s-* (20%) or a linking *-e(n)-* (11%). This distribution is different for compounds in which a derived noun appears as left constituent (17% of all compounds). Derived left constituents almost always occur with a linking element (*-s-*: 62.7%; *-en-*: 32.8%; *-∅-*: 4.6%). Linking elements are thus typical for derived forms, although they also occur with other left constituents. In addition to this general preference of derived forms to occur with linking elements, specific suffixes also tend to occur with particular linking elements. For instance, the diminutive suffix *-tje* and its allomorphs are always followed by the linking *-s-*. The suffix *-heid* (similar to English *-ness*), even though it appears with all three linking possibilities (*-s-*, *-en-*, and *-∅-*), evidences a very strong preference: 99% of all such compounds select *-s-*. These strong restrictions are also effective in the case of novel compounds (Krott, Baayen, & Schreuder, 2001, also chapter 2; Krott, Schreuder, & Baayen, in press, also chapter 3).

Historically, Dutch linking elements were case endings of the left constituent (Booij, 1996; Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997). Synchronically, they are homophonous with the two plural suffixes. However, they do not simply mark plural semantics. Note that the linking *-s-* may appear after left constituents that take a different plural suffix (e.g., sg. *geluid* 'noise', pl. *geluiden*, but *geluid+s+signaal* 'acoustic signal'). In addition, combinations of left constituents and linking elements cannot always be interpreted as plural forms (e.g., *boer+en+zoon* farmer+EN+son, 'son of a farmer'). On the other hand, *-en-* can be interpreted as a plural marker in many compounds and there is even evidence that it activates plural semantics in comprehension (Schreuder, Neijt, Van der Weide, & Baayen, 1998). This suggests that *-en-* and *-s-* do not serve just a single function.

A function for linking elements that has recently been proposed for German by Aronoff & Fuhrhop (submitted) is the opening function of closing suffixes. Closing suffixes are suffixes that are never followed directly by another suffix or stem. Aronoff and Fuhrhop point out that these German suffixes can appear in compounds, but that in such cases they are followed by a linking element. This suggests that the linking elements in German may have the function of making derived forms available for further word formation which otherwise have a morphological valency

of zero. Interestingly, German suffixes that are not closing suffixes are never followed by linking elements.

Dutch also has closing suffixes. Among the list that has been proposed by Booij & Baayen (in preparation) on quantitative grounds, the following prototypical closing suffixes appear in Dutch compounds: *-er*, *-erij*, *-heid*, *-ing*, *-iteit*, and *-ster*. However, as we will see below, in contrast to German, Dutch closing suffixes are not the only suffixes that are followed by linking elements (e.g., *-schap* in *leiderschap+s+stijl* 'leadership style').

The issue that we will address in this study is the functionality of Dutch linking elements when they follow derived forms. In particular, we will focus on the question whether they also have the function of opening derived forms that end in closing suffixes for further word formation. We address this question by means of a statistical analysis of the distributional properties of compounds and their constituents as attested in the CELEX lexical database.

Suffixes and their degree of overrepresentation

The distributional properties of compounds that we will focus on in this paper have been addressed in a previous study by Krott, Schreuder, & Baayen (1999, also chapter 7). They have focused on complex words that themselves contain complex constituents, and showed that words with different morphological structure are non-uniformly distributed as constituents in complex words. Their distribution significantly deviates from the distribution that one would expect under chance conditions. While monomorphemic words occur much more often than expected under chance conditions, i.e they are overrepresented, compounds appear much less than under chance conditions, i.e. they are underrepresented. Derived nouns turn out to be slightly, but significantly overrepresented. Krott et al. (1999, also chapter 7) also revealed that the degree of over- and underrepresentation correlates with word frequency and degree of productivity: Morphological categories containing many high frequency words are overrepresented, while categories with many low frequency words are underrepresented. In addition, suffixes that give rise to frequent derived words are typically overrepresented, while suffixes that give rise to infrequent derived words are underrepresented. In addition, highly productive suffixes are underrepresented, while unproductive suffixes are overrepresented. All these correlations hold both for left and right constituents of Dutch compounds.

The study by Krott et al. (1999, also chapter 7) did not consider any potential dif-

ferences in overrepresentation for compounds with different linking elements. We therefore group Dutch compounds according to the embedded linking element, i.e. *-en-*, *-s-*, and *-∅-*. In the following, we focus on the overrepresentation of derived forms as well as the correlations of the degree of overrepresentation of a derivation class with its frequency and productivity. In particular, we will investigate the degree of overrepresentation of the closing suffixes identified by Booij & Baayen (in preparation). If linking elements indeed have the function of opening derived forms ending in closing suffixes, one would expect that closing suffixes are especially overrepresented in compounds that contain linking elements. Thus, we interpret overrepresentation as an indication for the function of opening the preceding suffix for further word formation.

The linking -∅-

We calculate the degree of overrepresentation as in Krott et al. (1999, also chapter 7). The first column of Table 8.1 in the Appendix lists all types of derived words that occur in CELEX as left constituents of compounds without a linking element, i.e. *-∅-*. The column f lists the number of compounds for each derivation class. The total number of compounds, including compounds with non-derived left constituents, that contain no linking element is 3036. There seem to be only a few suffixes that occur frequently in these compounds (*-atie*, *-ie*, *-er*, *-ing*, *-tie*, *-te*, *-age*). However, in order to determine the over- or underrepresentation of a derivation class, we have to ascertain the expected number of compounds. To do so, we make use of the binomial model and estimate the probability to find a formation of a derivation class on the basis of all possible nominal constituents and the number of existing derived words of the derivation class in question. For instance, out of the possible 54403 nouns in CELEX, 373 nouns end in *-atie*, which means that the probability p to find a formation in *-atie* as a compound constituent is 373/54403 = .00686 (see column p of Table 8.1). Because there are 3036 different compounds without a linking element, the expected number of compounds ending in *-atie* among the compounds without a linking element is .00686 $*$ 3036 = 20.82. That is far less than the observed 114 compounds. Column E lists the expected numbers of types for all derivation classes. In order to determine whether the difference between the observed number and expected number is significant, we approximate the binomial model by a normal model and calculate Z-scores (Z = (f-np)/s), with s = $\sqrt{np(1-p)}$). Z-scores and corresponding standard deviations are listed in columns Z and s respectively. A positive Z-score indicates overrepresentation, while

a negative Z-score indicates underrepresentation. The column $sign$ lists the corresponding Bonferroni-adjusted significance levels.

In the following analyses, we only include suffixes that do occur with a particular linking possibility. The zero frequency of suffixes that do not appear is too inaccurate. It is possible that such a suffix shall never occur without a linking element or that it does occur, but only in a much bigger corpus. In addition, in the case of a zero frequency, the calculation of the Z-scores would be based only on the observations of these suffixes outside compounds.

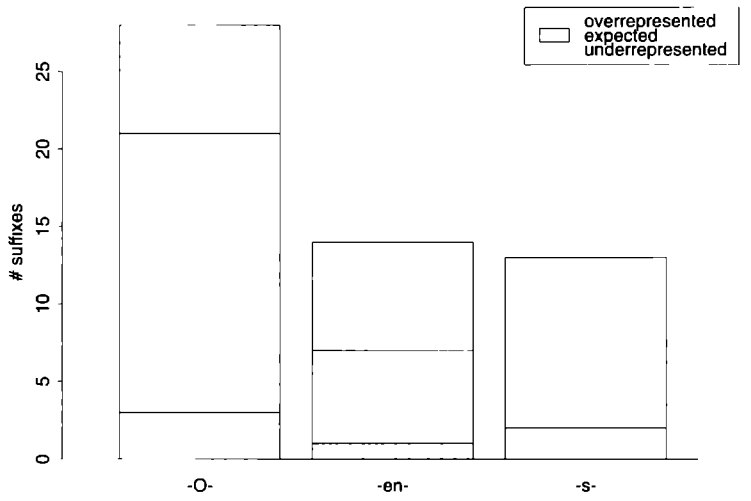


Figure 8.1: Number of suffixes that are overrepresented, underrepresented, or expected under chance condition for the linking possibilities *-en-*, *-s-*, and *-∅-*.

As both Table 8.1 and Figure 8.1 show, only roughly a third of all suffixes that appear in compounds without linking elements are either over- or underrepresented (10 out of 28). Most of them occur as expected given the number of derivations that exist in the language. Although there are more suffixes significantly overrepresented (7) than underrepresented (3), this difference is not significant (proportions test: p = .295). The closing suffixes are either highly underrepresented (*-er*, *-heid*, and *-ing*), or their number is as expected under chance conditions (*-ster* and *-erij*). The closing suffix *-iteit* does not appear in this group of compounds at all. Interestingly, closing suffixes are the only underrepresented suffixes in compounds without linking elements. Their underrepresentation is in line with the hypothesis that clos-

ing suffixes are followed by a linking element in order to be used as a left compound constituent.

As a next step, we investigate whether the degree of overrepresentation is correlated with the frequency and productivity of a derivation class. Krott et al. (1999, also chapter 7) have shown that such correlations exist for compounds, without distinguishing between compounds with different linking elements. The column labeled $fmean$ in Table 8.1 lists the mean log frequency of the derived forms for each derivation class in a corpus of 42 million wordforms available in the CELEX database. The upper left panel of Figure 8.2 shows that we indeed have a positive correlation between mean log frequency and Z-score, the measurement of overrepresentation (Pearson: r = .40, t(26) = 2.21, p = .037; Spearman: r_s = .64, p $<$ .001). The solid line represents the corresponding mean squares regression line. Closing suffixes are written in upper case letters, while all other suffixes are written in lower case letters. Recall that positive Z-scores indicate overrepresentation, while negative Z-scores indicate underrepresentation. The upper left panel of Figure 8.2 shows that derived forms ending in closing suffixes are of lower frequency and, as already mentioned, they are all underrepresented. The upper right panel of Figure 8.2 shows the correlation between degree of overrepresentation (Z-score) and the category-conditioned degree of productivity as defined in Baayen (1992). Note that a linear model is obviously inappropriate for the data, as can be seen from the dashed line, representing a non-parametric regression smoother (see Cleveland, 1979). We therefore tested the correlation with the means of a Spearman correlation analysis which shows that correlation is reliable (r_s = -.69,p $<$ .001). Summing up, these correlation analyses reveal that, in the case of compounds that do not contain linking elements, the degree of overrepresentation is reliably correlated with both frequency and productivity. These findings are comparable with the findings for compounds as a whole reported in Krott et al. (1999, also chapter 7).

The linking -en- and -s-

Table 8.2 summarizes the data for the calculation of the degree of overrepresentation for the derivation classes that occur as left constituents in compounds with the linking *-en-*. Only derivations in *-heid* are underrepresented, seven other classes are overrepresented. Only three of the six closing suffixes occur in compounds with *-en-*. Of these, the suffix *-heid* is significantly underrepresented and both *-iteit* and *-ing* are neither over- nor underrepresented. We therefore conclude that there is no clear evidence that *-en-* has the function of opening derived forms ending in the

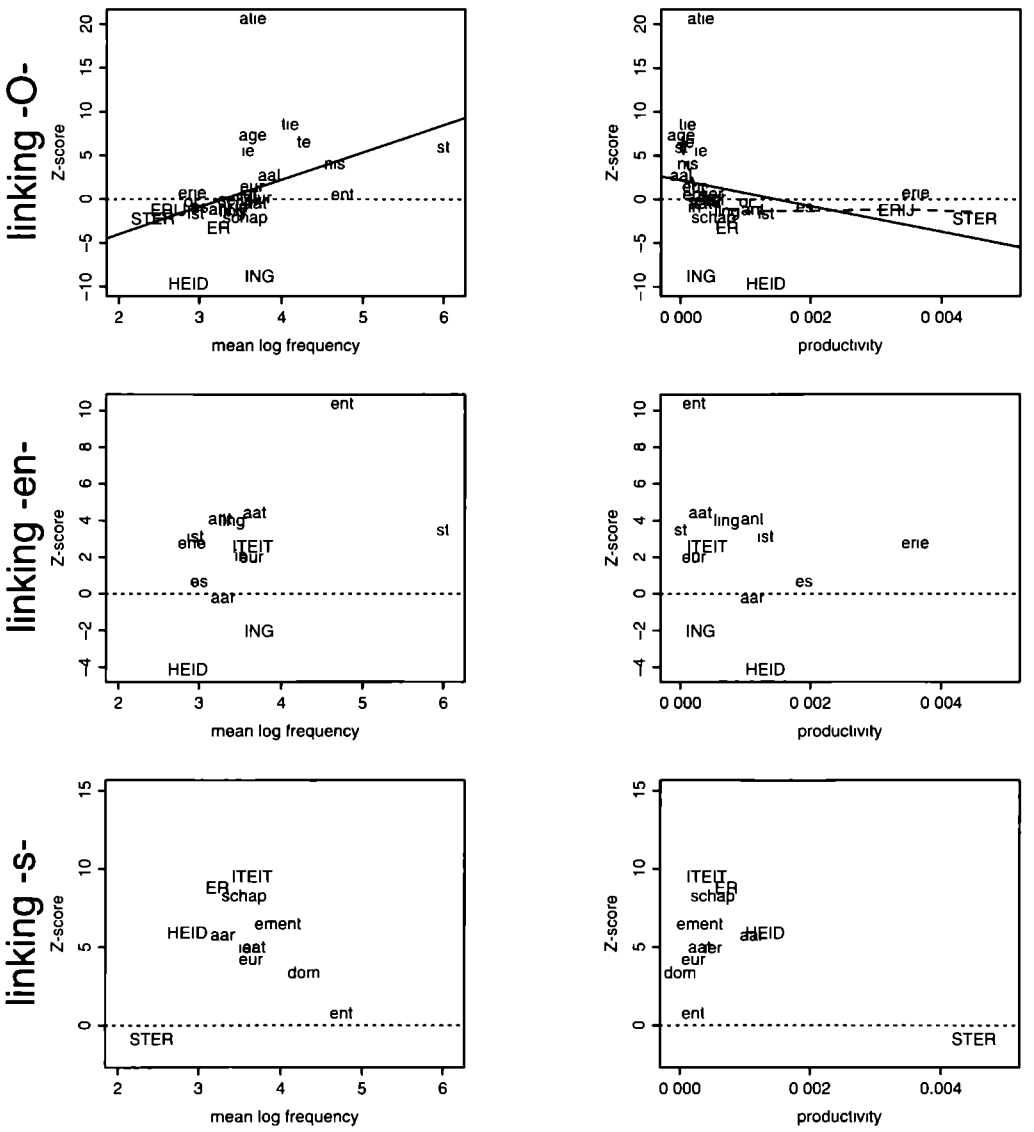


Figure 8.2: Derived forms as left constituents and correlations of their degree of overrepresentation (Z-score) with frequency and productivity in compounds containing the linking *-en-*, *-s-*, or *-∅-* (closing suffixes in upper case letters). The solid lines are mean squares regression lines, while the dashed lines represent a non-parametric regression smoother. The suffix *-ing* is not shown in the lower two panels because it lies outside the range.

six closing suffixes that were identified by Booij & Baayen (in preparation). Let us, however, assume that the opening function is not restricted to these six suffixes, but that linking elements might open also other suffixes for further word formation. That would mean that suffixes that are overrepresented with a linking element should not be overrepresented in compounds without linking elements or with *-s-*. This is indeed the case for the suffixes that are overrepresented with *-en-*, except for *-aat* that is also overrepresented with *-s-* (see below). Thus, although the linking *-en-* does not open the proposed prototypical closing suffixes for further compounding, it may have the less strong function of enhancing word formation for derived classes in general. This is also evident from Figure 8.1 that shows an increase in general overrepresentation (7/28=25% versus 7/14=50%) and a small decrease in underrepresentation (3/28=11% versus 1/14=7%) compared to compounds without linking elements, which is, however, not significant (Fisher: p=0.2663, two-tailed).

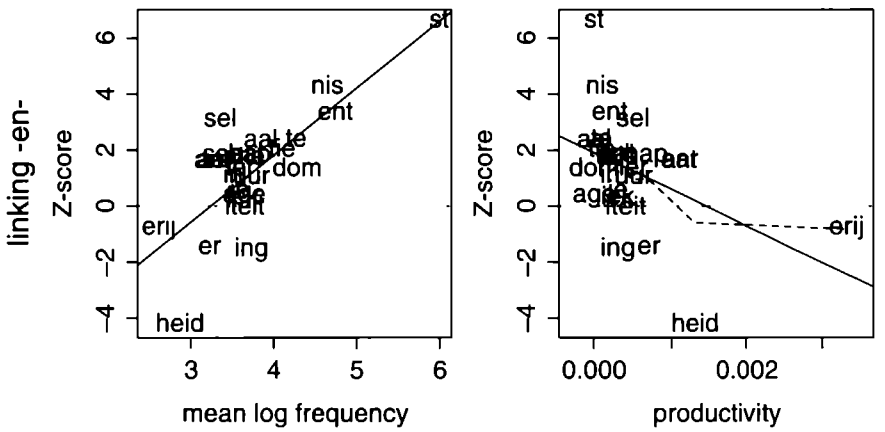


Figure 8.3: Derived forms in right constituents and correlations of their degree of overrepresentation (Z-score) with frequency and productivity in compounds containing the linking *-en-*. The solid lines are mean squares regression lines, while the dashed line represents a non-parametric regression smoother.

The middle left panel of Figure 8.2 shows the correlation between the degree of overrepresentation and mean log frequency for left derivation classes that occur with *-en-*. Surprisingly, this correlation is neither significant in a Pearson correlation analysis ($r = 0.45$, $t(12) = 1.72$, $p = .111$) nor in a Spearman analysis ($r_s = .35$,

p = .211). In addition, the correlation between overrepresentation and productivity is not significant either (r = -.20, t(12) = -.69, p = .502; r_s = -.27, p = .322). Apparently, derivation classes with high frequency members are not more likely to occur as left constituents in compounds with the linking element *-en-* than derivation classes with low frequency members. And it is also not the case that nouns ending in unproductive suffixes are used more often as left constituents than nouns ending in productive suffixes. The pattern of over- and underrepresentation of derivation classes as left constituents in compounds with *-en-* must be caused by another factor. Interestingly, these correlations are not significant only for left constituents.

Figure 8.3 shows how the degree of overrepresentation is correlated with mean log frequency as well as with productivity in the case of right constituents that are following the linking *-en-*. Both correlations are significant (frequency: r = .77, t(22) = 5.71, p < .001; r_s = .58, p = .006; productivity: r = -.43, t(22) = -2.25, p = .035; r_s = -.47,p = .024). The dashed non-parametric regression line in the right panel shows that a linear model is a reasonable description of the main trend in the data, except for the outlier *-erij*.

Table 8.3 summarizes the derivation classes and values of their overrepresentation in compounds with the linking *-s-*. There is no suffix that is underrepresented. As Figure 8.1 shows, almost all suffixes (11/13) are overrepresented. We now observe a significant increase in overrepresentation and a decrease in underrepresentation compared to compounds with *-en-* and *-∅-* (*-en-*: 7/14=50% overrepresented, 1/14=7% underrepresented; *-∅-*: 7/28=25% overrepresented, 3/28=11% underrepresented; Fisher test: p=.007, two-tailed). In addition, half of the overrepresented suffixes that do not belong to the set of the six proposed closing suffixes are neither over- nor underrepresented in compounds without linking elements. They are thus only overrepresented with *-s-*. Furthermore, five of the six closing suffixes appear, of which four are highly overrepresented. The closing suffix *-ster* is neither over- nor underrepresented, and *-erij* does not occur with *-s-*. The latter only occurs in compounds without any linking element, where it is underrepresented. The general increase in overrepresentation and the strong overrepresentation of the prototypical closing suffixes indicate that the linking element *-s-* does indeed open derived forms for further word formation, in particular so for those ending in closing suffixes.

The lower left panel of Figure 8.2 shows the correlation between the degree of overrepresentation and frequency for derivation classes in compounds that contain the linking *-s-*. This correlation is not reliable (r = .06, t(11) = .21, p = .840;

r_s = -.055, p = .842). The same holds for the correlation between the degree of overrepresentation and productivity (r = -.22, t(11) = -.76, p = .466; r_s = .16, p = .591), which is shown in the lower right panel of Figure 8.2. The suffix *-ing* has an extreme high Z-score which lies outside the range of the other suffixes. Therefore, *-ing* is not shown in the lower panels. As in the case of compounds with linking *-en-*, all these correlations are fully reliable in the case of derived classes that occur as right constituents (frequency: r = .66, t(28) = 4.65, p < .001; r_s = .68, p < .001; productivity: r = -.49, t(28) = -2.87, p = .006; r_s = -.60, p < .001). In other words, the only cases in which both kinds of correlations are not reliable concern left constituents in compounds that contain linking elements.

General discussion

In this study, we have focused on the function of Dutch linking elements that follow derived wordforms, in particular on the function of opening derived nouns that end in closing suffixes for compounding. We have addressed this issue by means of analyzing the degree of overrepresentation of derivation classes that occur as constituents in compounds with different linking elements. We have seen that the degree of overrepresentation varies significantly with the linking element in the compound. The prototypical closing suffixes *-er*, *-heid*, *-ing*, and *-iteit* are overrepresented with the linking *-s-*, while they hardly occur in compounds without linking elements or with the linking *-en-*. The suffixes *-ster* and *-erij* are both closing suffixes that are neither underrepresented nor overrepresented in all three kinds of compounds. These findings support the hypothesis that the linking *-s-* has indeed the function of opening closing suffixes for further word formation. There is no such evidence, however, for the linking *-en-*.

In the case of suffixes that are not prototypical closing suffixes (Booij & Baayen, in preparation), we observed a decrease in underrepresentation and an increase in overrepresentation in compounds with linking elements compared to compounds without linking elements. This difference is most prominent when comparing compounds with *-s-* and *-∅-*. Thus, the linking *-s-* also opens other suffixes than the closing suffixes for further word formation. The same holds for *-en-*, which reveals more over- and less underrepresentation than compounds without linking elements. Therefore, although *-en-* does not open prototypical closing suffixes, it opens other suffixes for further word formation. The main function of *-en-*, however, which also holds for left constituents that are not derived nouns, is the function of marking plu-

rality of the left constituent. Evidence for this function has been found in a previous perception study (Schreuder, Neijt, Van der Weide, & Baayen, 1999).

The question arises why opening closing suffixes with the means of a linking element is necessary or advantageous. From a processing point of view, there seems to be no a priori advantage in producing a compound with a linking element. From a perception point of view, however, linking elements cancel word boundaries that are indicated by closing suffixes. The word boundary is, however, only cancelled when the linking element is not the appropriate plural suffix. Interestingly, most of the prototypical closing suffixes take *-en* as their plural suffix, with the exceptions of *-ster* and *-er*. The linking *-s-* is therefore better suited to open closing suffixes.

The present study has also shown that linking elements break up the general distributional patterns for derived forms that occur as left compound constituents, while the distributional pattern for right derived constituent are still intact. The absence of correlational structure in the case of left constituents might be partly due to the function of opening closing suffixes. In addition, the overrepresentation of a derivation class in compounds with a particular linking element is probably induced by the strong restrictions on the combination of suffixes and linking elements. The overrepresentation might well arise in the course of the process of selecting linking elements for novel compounds, as described in Krott et al. (2001, also chapter 2) and Krott et al. (in press, also chapter 3). These studies presented evidence that linking elements are mainly selected on the basis of the distribution of linking elements in existing compounds that share the left constituent with the target compound, a set of compounds that we refer to as the left constituent family. In the case of derived left constituents, there is a clear effect of the distribution of linking elements in the compounds sharing the suffix with the left constituent of the target compound. However, there is further evidence that the analogical force of the left constituent family overrides the analogical force of the suffix (Krott et al., in press, also chapter 3). We think that the stronger analogical force of the left constituent family also contributes to disturbing the distributional pattern for derived left constituents.

Further evidence on support of this hypothesis is available in Krott et al. (2001, also chapter 2) who have shown that, in addition to a strong effect of the left constituent family, there is also an effect of the distribution of linking elements in the set of existing compounds sharing the right constituent with the target compound, a set that we refer to as the right constituent family. Interestingly, the analogical force of the right constituent family is much weaker than that of the left constituent family.

We suspect that it is too weak to affect the correlational patterns for right derived constituents.

In sum, the absence of the correlational pattern for frequency and productivity of the left constituent is probably due to a synergetic complex of factors including the opening function of the linking elements, the strong restrictions of suffixes on following linking elements, and the analogical force of the left constituent family.

References

Aronoff, M. and Fuhrhop, N.: submitted, Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint.

Baayen, R. H.: 1992, Quantitative aspects of morphological productivity, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1991*, Kluwer Academic Publishers, Dordrecht, pp. 109–149.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Booij, G. E.: 1996, Verbindingsklanken in samenstellingen en de nieuwe spellingregeling (Linking phonemes in compounds and the new spelling system), *Nederlandse Taalkunde* **2**, 126–134.

Booij, G. E. and Baayen, R. H.: in preparation, Suffix order in Dutch.

Cleveland, W. S.: 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Krott, A., Baayen, R. H. and Schreuder, R.: 2001, Analogy in morphology: modeling the choice of linking morphemes in Dutch, *Linguistics* **39**(1), 51–93.

Krott, A., Schreuder, R. and Baayen, R.: 1999, Complex words in complex words, *Linguistics* **37**(5), 905–926.

Krott, A., Schreuder, R. and Baayen, R. H.: in press, Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds, *in* R. Skousen (ed.), *Analogical Modeling: An Exemplar-Based Approach to Language*.

Schreuder, R., Neijt, A., Van der Weide, F. and Baayen, R. H.: 1998, Regular plurals in Dutch compounds: linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573.

Appendix

Table 8.1: **Derivation classes as left constituents in compounds with -$\emptyset$-**: f: number of types; $meanf$: mean log token frequency; $prod$: productivity; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); closing suffixes in upper case letters.

class	f	meanf	prod	fcel	p	E	s	Z	sign
-atie	114	3.64	0.0003	373	0.0069	20.8	4.6	20.5	**
-ie	38	3.58	0.0003	292	0.0054	16.3	4.0	5.4	**
-ER	25	3.22	0.0007	881	0.0162	49.2	7.0	-3.5	**
-ING	18	3.72	0.0003	1986	0.0365	110.8	10.3	-9.0	**
-tie	17	4.10	0.0001	51	0.0009	2.9	1.7	8.4	**
-te	16	4.27	0.0001	67	0.0012	3.7	1.9	6.3	**
-age	12	3.64	0.0000	35	0.0006	2.0	1.4	7.2	**
-st	8	5.99	0.0000	24	0.0004	1.3	1.2	5.8	**
-nis	7	4.64	0.0001	32	0.0006	1.8	1.3	3.9	**
-sel	6	3.35	0.0005	118	0.0022	6.6	2.6	-0.2	
-eur	5	3.63	0.0002	53	0.0010	3.0	1.7	1.2	
-ERIJ	5	2.61	0.0033	169	0.0031	9.4	3.1	-1.5	
-iek	4	3.62	0.0003	83	0.0015	4.6	2.2	-0.3	
-ier	3	3.59	0.0005	41	0.0008	2.3	1.5	0.5	
-ent	2	4.74	0.0002	28	0.0005	1.6	1.3	0.4	
-aal	2	3.84	0.0000	7	0.0001	0.4	0.6	2.6	
-aat	2	3.67	0.0003	56	0.0010	3.1	1.8	-0.6	
-ling	2	3.39	0.0007	92	0.0017	5.1	2.3	-1.4	
-or	2	2.90	0.0010	52	0.0010	2.9	1.7	-0.5	
-HEID	2	2.86	0.0013	1759	0.0323	98.2	9.8	-9.9	**
-uur	1	3.72	0.0005	20	0.0004	1.1	1.1	-0.1	
-schap	1	3.55	0.0005	113	0.0021	6.3	2.5	-2.1	
-in	1	3.50	0.0002	49	0.0009	2.7	1.7	-1.1	
-ant	1	3.25	0.0011	63	0.0012	3.5	1.9	-1.3	
-es	1	3.00	0.0019	53	0.0010	3.0	1.7	-1.1	
-ist	1	2.95	0.0013	90	0.0017	5.0	2.2	-1.8	
-erie	1	2.91	0.0036	10	0.0002	0.6	0.8	0.6	
-STER	1	2.42	0.0045	139	0.0026	7.8	2.8	-2.4	

Table 8.2: **Derivation classes as left constituents in compounds with -en-**: f: number of types; $meanf$: mean log token frequency; $prod$: productivity; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); closing suffixes in upper case letters.

class	f	meanf	prod	fcel	p	E	s	Z	sign
-ING	12	3.72	0.0003	1986	0.0365	21.7	4.6	-2.1	
-ent	6	4.74	0.0002	28	0.0005	0.3	0.6	10.3	**
-ITEIT	5	3.64	0.0004	159	0.0029	1.7	1.3	2.5	
-ling	5	3.39	0.0007	92	0.0017	1.0	1.0	4.0	**
-aat	4	3.67	0.0003	56	0.0010	0.6	0.8	4.3	**
-ant	4	3.25	0.0011	63	0.0012	0.7	0.8	4.0	**
-ist	4	2.95	0.0013	90	0.0017	1.0	1.0	3.1	*
-st	2	5.99	0.0000	24	0.0004	0.3	0.5	3.4	**
-eur	2	3.63	0.0002	53	0.0010	0.6	0.8	1.9	
-in	2	3.50	0.0002	49	0.0009	0.5	0.7	2.0	
-aar	1	3.29	0.0011	128	0.0024	1.4	1.2	-0.3	
-es	1	3.00	0.0019	53	0.0010	0.6	0.8	0.6	
-erie	1	2.91	0.0036	10	0.0002	0.1	0.3	2.7	*
-HEID	1	2.86	0.0013	1759	0.0323	19.2	4.3	-4.2	**

Table 8.3: **Derivation classes as left constituents in compounds with -s-**: f: number of types; $meanf$: mean log token frequency; $prod$: productivity; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level (*: .05; **: .01); closing suffixes in upper case letters.

class	f	meanf	prod	fcel	p	E	s	Z	sign
-ING	381	3.72	0.0003	1986	0.0365	36.7	5.9	58.0	**
-HEID	65	2.86	0.0013	1759	0.0323	32.5	5.6	5.8	**
-ER	51	3.22	0.0007	881	0.0162	16.3	4.0	8.7	**
-ITEIT	19	3.64	0.0004	159	0.0029	2.9	1.7	9.4	**
-schap	14	3.55	0.0005	113	0.0021	2.1	1.4	8.3	**
-aar	11	3.29	0.0011	128	0.0024	2.4	1.5	5.6	**
-aat	6	3.67	0.0003	56	0.0010	1.0	1.0	4.9	**
-ement	5	3.96	0.0003	27	0.0005	0.5	0.7	6.4	**
-eur	5	3.63	0.0002	53	0.0010	1.0	1.0	4.1	**
-ier	5	3.59	0.0005	41	0.0008	0.8	0.9	4.9	**
-dom	2	4.28	0.0000	15	0.0003	0.3	0.5	3.3	**
-ent	1	4.74	0.0002	28	0.0005	0.5	0.7	0.7	
-STER	1	2.42	0.0045	139	0.0026	2.6	1.6	-1.0	

Summary and conclusions

The main goal of this thesis has been to come to a better understanding of how speakers select linking elements for novel Dutch noun-noun compounds. We have seen that their selection cannot be predicted with a reasonable degree of accuracy on the basis of general rules, even though there is substantial agreement among Dutch speakers as to which linking element is appropriate. The rules that are proposed in the literature predict Dutch linking elements on the basis of phonological, morphological, and semantic properties of the initial constituent and the semantic relation between the two constituents (Van den Toorn, 1981a, 1981b, 1982a, 1982b; Mattens, 1984; Haeseryn, Romijn, Geerts, Rooij, & Van den Toorn, 1997; Booij & Van Santen, 1995; Booij, 1996). However, the phonological and morphological rules jointly apply to only 51% of the compounds listed in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), and the prediction accuracy for this subset of compounds is only 63%, which is 32% of all compounds.

This thesis presents a novel approach to Dutch linking elements and explains their selection on the basis of paradigmatic analogy to existing compounds. This paradigmatic notion of analogy is based on formal similarity measures that are calculated over stored exemplars in an instance base. It is therefore not the traditional notion of analogy that is used to explain incidental exceptional word formation on the basis of some perceived similarity to a single ad-hoc example. The main result of this thesis is the strong evidence it reveals that linking elements in Dutch compounds are analogically selected on the basis of the distribution of linking elements in the paradigmatic sets of stored compounds that share the left (or right) constituent with the target compound. I have referred to the latter as the left and right constituent family. Evidence for this hypothesis has been obtained by both experiments in which participants had to form novel compounds and by simulation studies of novel and existing compounds with exemplar-based models of analogy (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2000; Skousen, 1989).

Summary of results

Chapter 2 has presented two experiments that systematically examined the analogical prediction of Dutch linking elements in novel compounds. The aim of this chapter has been to ascertain whether the distribution of linking elements in the left and right constituent families determines the selection of linking elements in novel compounds. Two cloze tasks in which participants had to select the appropriate linking element for novel compounds investigated the analogical force of the constituent families by varying the proportions of the linking elements *-en-* and *-s-* in both the left and right constituent families. The results of the experiments revealed that the selection of Dutch linking elements can indeed be predicted with a high degree of accuracy on the basis of the distribution of linking elements in these left and right constituent families. Especially the left constituent family emerged as a strong factor. The literature on Dutch linking elements proposes rules that predict the occurrence of specific linking elements after particular suffixes (Van den Toorn, 1981a; 1981b). A further experiment therefore examined the predictive force of the final suffix of the left constituent by using derived pseudo-nouns as initial constituents. The results confirm an effect of the distribution of linking elements in the suffix family, i.e. the set of compounds sharing the suffix of the initial constituent with the target compound.

Simulation studies with the computational exemplar-based analogical model TiMBL (Daelemans et al., 2000), confirmed the analogical effect of the left constituent family. Of the roughly 32,000 noun-noun compounds listed in the CELEX lexical database (Baayen, Piepenbrock, & Gullikers, 1995), TiMBL correctly predicted 92.5% when trained on the left constituent. This is remarkable considering the prediction accuracy of 32% that was obtained by applying the phonological and morphological rules proposed in the literature. TiMBL also successfully predicted the choices of linking elements that were given by the participants in our experiments. Using the left constituent as the analogical basis, TiMBL correctly predicted 78.8% of the majority choices for the novel compounds in the experiment examining the linking *-en-* and even 87.8% of the majority choices in the case of the experiment examining the linking *-s-*. Apparently, the model and the participants find the task equally difficult since the model's predictions come very close to the average agreement of a participant with the majority choice (*-en-*: 85.1%; *-s-*: 83.5%).

Finally, this chapter also presented a first outline of a psycholinguistically plausible interactive activation model that captures the paradigmatic analogical effect of the constituent families. In this model, the semantic and syntactic representations

of the left and right constituents send activation to the lexeme representations of the compounds that are contained in the corresponding constituent families. Subsequently, the members of the constituent families send activation onwards to the linking elements that they contain. The linking element that receives the highest activation has the highest probability to be selected for the target compound.

Chapter 3 has addressed the question whether different analogical factors such as the rime family, the suffix family, and the left constituent family are equally strong or whether they are hierarchically ordered. I first established that in addition to the suffix family and the constituent families, the rime family of the left constituent, i.e. the set of compounds that share the rime of the initial constituent with the target compound, plays a role as well, again using a cloze-task with left pseudo-nouns. Compared to the experiments that used existing left nouns and derived pseudo-nouns, there was considerably more variation in the responses and participants reported this task as being very difficult. These observations already show that the effect of the rime is weak.

In three further cloze-task experiments, the effects of the left constituent family, the suffix family, and the rime family were compared with each other. In the first experiment, I used left derived constituents for which I expected different linking elements depending on whether the prediction is based on the constituent family or on the suffix family. In the second experiment, I similarly varied the prediction of the constituent family and the rime family, and in the third experiment, I varied the prediction of the suffix family and the rime family. The results revealed that the bias of the constituent family overrules the bias of the suffix family as well as the bias of the rime family, and that the bias of the suffix family overrules the bias of the rime family. I concluded that the analogical factors left constituent family, suffix family, and rime family are hierarchically ordered.

In addition to these experiments, this chapter also presented a comparison of two exemplar-based analogical models TiMBL (Daelemans et al., 2000) and AML (Skousen, 1989). Simulation studies of the existing Dutch compounds listed in CELEX as well as of the responses given in the above experiments showed that the prediction accuracies of the two models are very similar. The models did not differ either with respect to their prediction uncertainty.

These simulation studies confirmed the hierarchy of analogical factors which had been observed experimentally. Firstly, this hierarchy was mirrored in the information gain values of the three factors, a measure that estimates the relevance of a factor

for the choice of the linking element. Secondly, the models predicted the observed responses significantly better when their prediction was based on higher-ranked factors than when it was based on lower-ranked factors. In fact, the highest-ranked factor that was available in the input emerged to be crucial for obtaining a high prediction accuracy. I have offered two possible explanations for these results. First, only the highest-ranked information that is given in the input might be used to determine the analogical set. Second, all information that is provided by the input might be used. Lower-ranked information, however, would then be masked by the strong analogical force of the highest-ranked information.

Chapter 4 shifted the focus from factors that are based on the form of the constituents (rime family, suffix family, and constituent family) to possible semantic effects of the left and right constituents. Van den Toorn (1982a) mentions that semantic factors such as the semantic class of the left constituent or the semantic relation between the two constituents might affect the choice of linking elements. A statistical analysis of the constituent families of the compounds that had been used in the first two experiments of chapter 2 revealed that there is a relation between the semantic class of the left and right constituent and the choice of the linking element. The results of a cloze-task experiment confirmed that the animacy and concreteness of the left constituent significantly affect the choice of the linking element in a novel compound. There was, however, no evidence for such an effect of the right constituent. A post-hoc analysis of the experiments presented in chapter 2 confirmed these results. In all post-hoc analyses, the semantic effect of the left constituent turned out to be independent of the form effects of the left and right constituent families.

Chapter 5 has examined the effect of the left and right constituent family when linking elements for novel Dutch compounds have to be selected under time pressure, using a decision task in which participants had to press one of two push buttons depending on whether they selected the linking element *-en-* or another linking element. I replicated the analogical force of the left and right constituent family on the choice of linking elements. In addition, the left constituent family turned out to determine the response latencies, with a stronger bias leading to shorter latencies. Interestingly, the weaker analogical effect of the right constituent family on the choices was absent for the response latencies. I explained this dissociation by means of a two-stage selection process based on the interactive activation model

presented in chapter 2. In the first stage of the activation process, a linking element is selected on the basis of the initial activation received from the left and right constituent families. In the subsequent processing stages, activation flows back and forth between the constituent families and the linking elements until the selected linking element reaches an activation level that allows response execution. This activation resonance leads to a rapid increase of activation in the left constituent family and its linking elements, while the activation of the right constituent family increases slowly. A simulation study using an implementation of this interactive activation model provided excellent fits to the experimental data, both with respect to the choices of linking elements as well as with respect to response latencies.

The simulation studies of chapter 4 have suggested that AML and TiMBL are excellent tools for predicting the outcome of regular and irregular analogical processes. While it might be objected that these models are not psycholinguistically relevant, I have demonstrated in chapter 5 that both the analogical selection of Dutch linking elements and its time course can be modeled by means of an interactive activation model that is formally equivalent to TiMBL's IB1-IG algorithm.

Chapter 6 addressed the question whether the paradigmatic analogical effects of the left and right constituent families is present in another language that contains partly-predictable linking elements, namely German. Preliminary evidence for an analogical effect of constituent families has been reported by Dressler, Libben, Stark, Pons, & Jarema (2001). In their paper-and-pencil experiment, participants had to select the appropriate linking elements for novel German noun-noun compounds, grouped into ten categories according to properties such as gender and rime. Although the responses in this experiment mainly followed the expectations based on the rules for these categories, Dressler et al. also report some evidence indicating the possibility of an analogical effect of the left constituent family. Following the design of the Dutch experiments, the three experiments reported in chapter 6 revealed an analogical effect of the left constituent family for the German linking possibilities *-s-*, *-(e)n-*, and *-∅-*. In contrast to Dutch, however, these experiments did not yield any evidence for an effect of the right constituent family.

Simulation studies, using TiMBL, of the existing German compounds in CELEX as well as of the responses given by the participants in the experiments showed that not only the left constituent families but also the rime, gender, and inflectional class families of the initial constituent simultaneously affect the selection of German linking elements. In the case of existing compounds, TiMBL reaches the highest

prediction accuracy when the prediction is based on the combination of all these factors. However, in the case of the experiment that tested the linking *-(e)n-*, the prediction accuracy was higher when the analogical prediction was based on only the properties of the left constituent (rime, gender, and inflection class) and not on the left constituent family. Conversely, in the case of the *-s-* experiment, the left constituent family leads to the highest independent prediction accuracy, which cannot be enhanced any further by including other factors. Apparently, different factors are relevant for different subsets of compounds. The simulation studies with TiMBL show that a model for analogy can predict the effect of factors such as rime, gender, and inflectional class of the initial constituent. This shows that rules, based on a division of compounds into categories, are not necessary to explain the selection of German linking elements. I have also outlined how the different analogical factors can be captured by an interactive activation model. Future research will have to clarify whether an implementation of this model can correctly predict the selection of linking elements in all different kinds of German compounds.

Chapter 7 is a study on the lexical statistics of word formation that provides the analytical tools for the statistical study on the linking elements in chapter 8. Chapter 7 focused on the over- and underrepresentation of complex words in complex words. I observed that monomorphemic nouns are used more often as immediate constituents in compounds than one would expect on the basis of the proportion of monomorphemic nouns in the set of all Dutch nouns. In contrast to monomorphemic constituents, compounds are highly underrepresented as immediate constituents and derived nouns are slightly overrepresented. I also showed that the degree of overrepresentation is correlated with the frequency and length of the constituents. Frequent and short nouns are overrepresented, while infrequent and long nouns are underrepresented. In the case of derived nouns, the degree of overrepresentation is also correlated with the productivity of the suffix. Productive suffixes are underrepresented, while unproductive suffixes are overrepresented. This pattern of results suggests that higher frequency, shorter forms with less productive suffixes, which are likely to be stored as units in the mental lexicon (Hasher & Zacks, 1984; Sereno and Jongman, 1997; Schreuder, De Jong, Krott, & Baayen, 1999; Baayen, Schreuder, De Jong, & Krott, in press), are more easily available for further word formation than lower frequency, longer words with more productive suffixes.

Chapter 8 has called attention to the function of Dutch linking elements when they follow derived nouns. In particular, it has addressed the question whether linking elements open closing suffixes for further word formation, which has been suggested to be the function of linking elements in German (Aronoff & Fuhrhop, submitted). However, the prototypical Dutch closing suffixes that have been listed by Booij and Baayen (in preparation) occur both with and without linking elements. In addition, other non-closing Dutch suffixes are also followed by linking elements. Therefore, this study did not examine the presence of linking elements for different suffixes as such. Instead, it analyzed the degree of overrepresentation of suffixes in compounds as a function of the different linking elements in these compounds. I observed that suffixes tend to be more overrepresented and less underrepresented in compounds that contain *-s-* or *-en-* than in compounds that do not contain linking elements. Apparently, *-s-* and *-en-* open derived words for further word formation to some extent. This opening function does not appear to be the main function of *-en-*, however. Its main function may well be to mark the plurality of the left constituent (see Schreuder, Neijt, Van der Weide, & Baayen, 1999). By contrast, the predominant function of the linking element *-s-* might indeed be the opening function: *-s-* reveals a high degree of overrepresentation with prototypical closing suffixes that are otherwise highly underrepresented.

Interestingly, the correlational patterns between properties of compound constituents that have been observed in chapter 7, such as the degree of overrepresentation, the frequency, and the productivity of derived forms, were absent for left constituents of precisely those compounds that contain an overt linking element. The correlational patterns for right constituents of all kinds of compounds were not affected by the presence of linking elements. I explained these results in terms of a conspiracy of factors: the opening function of the linking elements, the preference of certain suffixes for particular linking elements, as observed in chapters 2 and 3, and the strong paradigmatic force of the left constituent family that overrules the effect of the suffix family. The joint effect of these factors conspire to mask the effects of properties of morphological categories such as frequency and productivity.

Dutch and German linking elements - a comparison

We have seen in chapter 6 that the systems of German and Dutch linking elements differ in complexity. German not only has a larger set of linking elements (*-s-*, *-e-*, *-n-*, *-en-*, *-ens-*, *-es-*, and *-er-* versus *-s-*, *-en-*, and *-e-* in Dutch), its linking elements

-e- and *-er-* can also trigger umlaut in a preceding umlautable vowel. This difference in complexity of the system of linking elements in both languages is mirrored in the difference in complexity of their inflectional systems. Interestingly, the inflectional system plays a much more important role for the selection of German linking elements than for the selection of Dutch linking elements.

First consider Dutch. In this language, the strongest factor determining the selection of linking elements is the left constituent family, complemented by weaker effects of the right constituent family and properties of the left constituent such as its rime, its suffix, or its semantic class. It emerges from the experiments that the left constituent, the right constituent, and the semantic class of the left constituent all contribute effectively to the selection of the linking element. The factors of the left constituent, the suffix, and the rime, by contrast, appear to be hierarchically ordered with a measurable effect for only the highest factor in the hierarchy.

Next, consider German. The German study revealed effects of the left constituent family and effects of properties of the left constituent such as rime, gender, and inflection class. Interestingly, in contrast to Dutch, there was no evidence for any effect of the right constituent family. Furthermore, the German simulation studies suggest that the left constituent family as well as properties of the left constituent all independently affect the selection at the same time. There is no trace of a hierarchical ordering of factors. Interestingly, inflectional properties play a much more important role for German linking elements than for Dutch linking elements, as witnessed by the effect of gender and inflection class.

German and Dutch linking elements do not only differ with respect to the complexity of the analogical systems driving their selection, but also with respect to their function. Aronoff and Fuhrhop report that German linking elements open closing suffixes for further word formation (Aronoff & Fuhrhop, submitted). In the case of Dutch linking elements, this opening function is clearly present for the linking *-s-*, while the main function of the linking *-en-* appears to be to mark the plurality of the left constituent, although *-en-* also occurs more often in combination with suffixes than expected under chance conditions. A question that requires further research is whether German linking elements have other functions in addition to the opening function, and whether they may also mark, as in Dutch, the plurality of the preceding constituent. In the case of Dutch, future research will have to clarigy whether an implemented interactive activation model can also provide higher prediction accuracies for Dutch linking elements when factors such as the rime, the suffix or the semantic class of the left constituent are included in the model.

Stress patterns

One possible factor that has not been studied yet in this thesis is the potential effect of the stress pattern on the choice of the linking elements. A general characteristics of stress is its tendency of being rhythmically distributed, leading to an alternating stress pattern (Hayes, 1995; for Dutch see Booij, 1995). Given this tendency, one might expect that the linking *-en-* is used to avoid stress clash. Note that only *-en-* adds a new syllable, the linking *-s-* leaves the stress pattern untouched. In other words, one could expect to observe *-en-* more often in compounds that otherwise would instantiate a stress clash. Krebbers (2000) presented experimental evidence that participants indeed use the linking *-en-* more often in the case of a potential stress clash in compounds constructed out of pseudo-words. In order to better understand the role of stress in compounds with existing constituents, we conducted two studies: a statistical study of the CELEX compounds and a simulation study of these compounds with TiMBL.

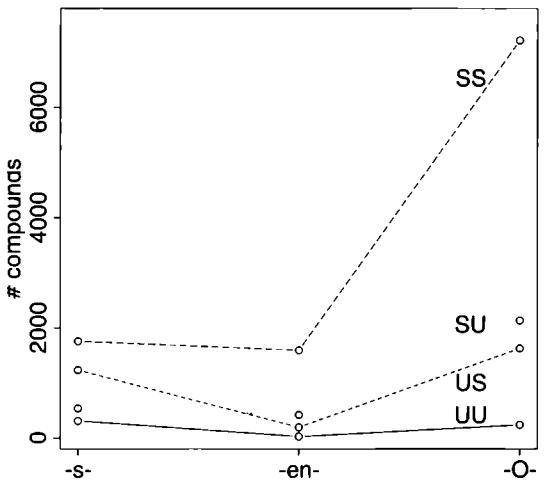


Figure 9.1: Number of compounds with the three linking possibilities *-s-*, *-en-*, and *-∅-* embedded in varying stress patterns (SS: stressed-stressed; US: unstressed-stressed; SU: stressed-unstress; UU: unstressed-unstressed).

For both studies, we selected the 12537 noun-noun compounds listed in CELEX in which a linking *-en-* is possible, i.e. compounds with left constituents that take *-en* as their plural suffix. We enriched the stress patterns given by CELEX by marking secondary stress on the last syllable of the left constituent and the initial syl-

lable of the right constituent. Figure 9.1 shows for the three linking possibilities *-s-*, *-en-*, and *-Ø-* the number of compounds with the four possible stress patterns stressed-stressed (SS), stressed-unstressed (SU), unstressed-stressed (US), and unstressed-unstressed (UU). A logit analysis of the observed number of compounds revealed main effects of the adjacent left stress ($F(1,2) = 422.3$, $p = .002$), the adjacent right stress ($F(1,2) = 453.1$, $p = .002$), and the linking element ($F(2,2) = 285.1$, $p = .003$). There is also a significant interaction between left stress and linking element ($F(2,2) = 48.3$, $p = .020$) and no interactions between left and right stress ($F(1,2) = 5.3$, $p = .147$), nor between right stress and linking element ($F(2,2) < 1$). The main effects of the left and right stress simply show that the different stress patterns are not equally frequent. As can be seen in Figure 9.1, compounds containing a stress clash are very common, while compounds with two unstressed syllables that immediately follow each other are uncommon. The interaction between left stress and linking element is more relevant for our question. Following a stressed syllable, the linking *-en-* (2015) and the linking *-s-* (2299) are similarly frequent, but no overt linking element is the most common choice. Following an unstressed syllable, *-en-* (227) occurs less often than *-s-* (1549) or no linking element (1872). Most importantly, the linking *-en-* is not the predominant linking element in compounds with a potential stress clash. The strong effect of the constituent family probably overrules the tendency of an alternating stress pattern. However, the significant interaction of the left stress and the linking element implies that the left stress may be relevant for the occurrence of linking elements in Dutch compounds.

We therefore ran simulation studies with TiMBL to ascertain whether stress contributes independently to the selection of linking elements in Dutch. Training the model on just the left constituent of the 12537 compounds mentioned above allows TiMBL to reach a prediction accuracy of 90.4%. If the model is trained on the left constituent, the rime of the left constituent, and the suffix of the left constituent, the prediction accuracy increases significantly to 91.1% (proportions test: $p = .055$). Including also the stress of the left syllable does not increase the prediction accuracy. The left stress by itself correctly predicts only 63.7% of the compound forms. Including the right stress into the training set with or without the left stress also never changes the prediction accuracy. We therefore conclude that the factor stress does not reliably affect the occurrence of linking elements, at least not in existing compounds containing a left constituent that takes an *-en* plural suffix.

Implications

The results reported in this thesis have important implications for the on-going debate between the single-route approach and the dual-route approach on the processing of regular and irregular morphological forms. The dual-route model (e.g., Anshen & Aronoff, 1988; Pinker, 1991, 1997, 1999; Pinker & Prince, 1991; Marcus, Brinkman, Clahsen, Wiese, & Pinker, 1995; Clahsen, 1999), which represents the traditional linguistic approach, assumes that regular and irregular morphological formations are handled by two distinct mechanisms. Novel regular forms are built by means of rules, while novel irregular formations are built by analogy to some stored exemplar. In this approach, rules are the only productive means for word formation. Novel irregular formations are taken to be rare and exceptional. In contrast to this dual-route approach, single-route approaches (e.g., Rumelhart & McClelland, 1986; Plunkett & Juola, 1999; Rueckl, Mikolinski, Raveh, Miner, & Mars, 1997; Seidenberg & Gonnerman, 2000; Skousen, 1989; Daelemans et al., 2000) assume that regular and irregular complex words are handled by one and the same mechanism.

Dutch linking elements form an interesting testing ground for these two contrasting approaches, since the choice of linking elements is only partly-predictable by standard symbolic rules, while it is fully productive with surprisingly high agreement among speakers. A traditional dual-route approach would have to explain the linking elements of many compounds (about 70% in our experiments) by appealing to exceptional analogy, which is less than satisfactory for such a productive word formation process. As we have seen in this thesis, in contrast to dual-route models, analogical models that base their prediction on stored exemplars successfully capture the selection of Dutch linking elements. Therefore, analogical models provide a fuller and more accurate account of the use of Dutch linking elements.

References

Anshen, F. and Aronoff, M.: 1988, Producing morphologically complex words, *Linguistics* **26**, 641–655.

Aronoff, M. and Fuhrhop, N.: submitted, Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Schreuder, R., De Jong, N. H. and Krott, A.: in press, Dutch inflection: the rules that prove the exception, *in* S. Nooteboom, F. Weerman and F. Wijnen (eds), *Storage and Computation in the Language Faculty*, Kluwer Academic Publishers, Dordrecht.

Booij, G.: 1995, *The phonology of Dutch*, Clerendon Press, Oxford.

Booij, G. and Van Santen, A.: 1995, *Morfologie. De Woordstructuur van het Nederlands* (Morphology. The Structure of Dutch Words), Amsterdam University Press, Amsterdam.

Booij, G. E.: 1996, Verbindingsklanken in samenstellingen en de nieuwe spellingregeling (Linking phonemes in compounds and the new spelling system), *Nederlandse Taalkunde* **2**, 126–134.

Booij, G. E. and Baayen, R. H.: in preparation, Suffix order in Dutch.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 2000, TiMBL: Tilburg Memory Based Learner Reference Guide. Version 3.0, *Report ILK 00-01*, Computational Linguistics Tilburg University.

Dressler, W. U., Libben, G., Stark, J., Pons, C. and Jarema, G.: 2001, The processing of interfixed German compounds, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp. 185–220.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and Van den Toorn, M.: 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen.

Hasher, L. and Zacks, R. T.: 1984, Automatic processing of fundamental information. the case of frequency of occurrence, *American Psychologist* **39**, 1372–1388.

Hayes, B.: 1995, *Metrical Stress Theory: Principles and Case Studies*, University

of Chicago Press, Chicago.

Krebbers, L.: 2000, *Olifantsoep Of Olifantensoep? Het Gebruik van de Tussenklank Bekeken bij Taalverwerving en Nonsenswoorden (Elephant Soup or Elephants Soup? The Use of Linking Elements in Acquisition and Nonce Words)*, Master's thesis, University of Nijmegen.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R. and Pinker, S.: 1995, German inflection: The exception that proves the rule, *Cognitive Psychology* **29**, 189–256.

Mattens, W. H. M.: 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343.

Pinker, S.: 1991, Rules of language, *Science* **153**, 530–535.

Pinker, S.: 1997, Words and rules in the human brain, *Nature* **387**, 547–548.

Pinker, S.: 1999, *Words and Rules: The Ingredients of Language*, Weidenfeld and Nicolson, London.

Pinker, S. and Prince, A.: 1991, Regular and irregular morphology and the psychological status of rules of grammar, *Proceedings of the 1991 meeting of the Berkeley Linguistics Society.*

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S. and Mars, F.: 1997, Morphological priming, fragment completion, and connectionist networks, *Journal of Memory and Language* **36**(3), 382–405.

Rumelhart, D. E. and McClelland, J. L. (eds): 1986, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, MIT Press, Cambridge, Mass.

Schreuder, R., De Jong, N. H., Krott, A. and Baayen, R. H.: 1999, Rules and rote: beyond the linguistic either-or fallacy, *Behavioral and Brain Sciences* **22**, 1038–1039.

Schreuder, R., Neijt, A., Van der Weide, F. and Baayen, R. H.: 1998, Regular plurals in Dutch compounds: linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Sereno, J. and Jongman, A.: 1997, Processing of English inflectional morphology, *Memory and Cognition* **25**, 425–437.

Skousen, R.: 1989, *Analogical Modeling of Language*, Kluwer, Dordrecht.

Van den Toorn, M. C.: 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is (Linking phonemes in compounds with derived forms as first constituents), *De Nieuwe Taalgids* **74**, 197–205.

Van den Toorn, M. C.: 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is (Linking phonemes in compounds with loanwords as first constituents), *De Nieuwe Taalgids* **74**, 547–552.

Van den Toorn, M. C.: 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I)), *De Nieuwe Taalgids* **75**(1), 24–33.

Van den Toorn, M. C.: 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for the regulation of linking phonemes in nominal compounds II), *De Nieuwe Taalgids* **75**(2), 153–160.

Samenvatting

Het onderzoek in deze dissertatie richt zich op een produktief woordvormingsproces in de Nederlandse taal, het creëren van een nieuwe samenstelling door de combinatie van twee zelfstandige naamwoorden met een tussenklank. Ruim een derde deel van de bestaande Nederlandse samenstellingen (35% in de lexikale databank CELEX, zie Baayen, Piepenbrock, & Gulikers, 1995) bevat de tussenklanken *-s-* (*schaap+s+kooi*, CELEX: 25%) of de tussenklank *-en-* (*boek+en+kast*) dan wel de orthografische variante *-e-* (*zonn+e+schijn*, CELEX: samen 11%). Standaard linguïstische analyses beschrijven het voorkomen van Nederlandse tussenklanken met behulp van fonologische, morfologische, en semantische regels (bijvoorbeeld Van den Toorn, 1981a, 1981b, 1982a, 1982b; Mattens, 1984; Haeseryn, Romein, Geerts, Rooij, & Van den Toorn, 1997; Booij en Van Santen, 1995; Booij, 1996). Deze regels zijn echter, zoals Van den Toorn (1982a) opmerkt, niet meer dan tendensen. Dat het inderdaad gaat om tendensen blijkt uit het feit dat de fonologische en morfologische regels samen alleen op 51% van de samenstellingen in CELEX van toepassing zijn en dat zij 63% van de tussenklanken in deze subgroep voorspellen, wat neerkomt op een successcore van slechts 32% van alle samenstellingen in CELEX. Dit onderzoek laat zien dat de voorspellingsnauwkeurigheid van deze regels ook onbevredigend is voor tussenklanken die proefpersonen voor nieuwe samenstellingen kiezen.

Deze dissertatie presenteert een nieuwe benadering die de keuze van Nederlandse samenstellingen op basis van paradigmatische analogie met bestaande samenstellingen verklaart. Deze paradigmatische notie van analogie is gebaseerd op een formele definitie die overeenkomsten tussen een doelsamenstelling en opgeslagen voorbeelden in een databank bepaalt. Deze notie van analogie is dus niet identiek met de traditionele notie die toevallige exceptionele woordvorming verklaart op basis van een herkenbare overeenkomst met een klein aantal ad-hoc voorbeelden. Het belangrijkste resultaat van deze dissertatie is de sterke evidentie dat tussenklanken in Nederlandse samenstellingen analogisch geselecteerd

worden. Zij volgen de distributie van tussenklanken in paradigmatische groepen van opgeslagen samenstellingen die de linker (of rechter) constituent met de doelsamenstelling delen. Deze paradigmatische groepen worden rechter en linker constituentfamilies genoemd.

Dit proefschrift is als volgt georganiseerd. Na een introductie van de vraagstellingen in hoofdstuk 1, presenteert hoofdstuk 2 een eerste serie van produktie-experimenten waarin proefpersonen uit twee zelfstandige naamwoorden nieuwe samenstellingen moeten vormen. Daarbij mogen zij, wanneer nodig, tussenklanken gebruiken. Deze experimenten tonen duidelijk aan dat de keuze van tussenklanken in nieuwe samenstellingen analogisch gedetermineerd wordt door de distributie van tussenklanken in zowel de linker als de rechter constituentfamilies, en dat het finale suffix van de linker constituent (de suffix familie) ook de keuze beïnvloedt. Simulatiestudies van deze resultaten met TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2000), een computationeel model voor analogie, ondersteunen de rol van de constituentfamilie als de primaire basis voor analogische voorspelling. Deze voorspellingen overtreffen de voorspellingen op basis van de regels uit de literatuur in hoge mate. Vervolgens presenteert dit hoofdstuk een psycholinguïstisch model dat de genoemde non-deterministisch vormselectie modelleert zonder symbolische representatie op te geven zoals in connectionistische modellen (bijvoorbeeld, Plunkett & Juola, 1999; Rueckl, Raveh, Miner, & Mars, 1997; Seidenberg & Gonnerman, 2000).

Hoofdstuk 3 presenteert evidentie voor een andere analogische factor, de rijm van de linker constituent. Gegeven deze resultaten komt de vraag op welk van deze factoren de sterkste invloed heeft. Daarop geeft een reeks van produktie-experimenten antwoord. Deze experimenten laten een hiërarchische volgorde van de linker constituent, het suffix, en de rijm zien. Dat wil zeggen dat de distributie van de tussenklanken in de constituentfamilies blijkbaar het sterkste effect op de keuze heeft. Dit effect overheerst de suffix- en rijmeffecten, terwijl het suffix-effect het rijmeffect overheerst. Het is nog niet duidelijk of deze factoren parallele invloed hebben. Het is mogelijk dat de hoogste factor in de hiërarchie alle andere factoren uitschakelt. De resultaten van deze produktie-experimenten worden vervolgens met twee verschillende modellen van analogie gemodelleerd: TiMBL en AML (Skousen, 1989). Beide modellen halen een vergelijkbare voorspellingsaccuraatheid zowel voor tussenklanken in de nieuwe samenstellingen die in de experimenten zijn gebruikt, als voor bestaande samenstellingen in CELEX.

Hoofdstuk 4 verlegt de focus van factoren die op de vorm van de constituenten

gebaseerd zijn (rijmfamilie, suffixfamilie, en constituentfamilie) naar mogelijke semantische effecten van de linker en rechter constituenten. Een statistische analyse van de constituentfamilies van de samenstellingen die in de experimenten van hoofdstuk 2 zijn gebruikt laat zien dat er een relatie is tussen de semantische klasse van de linker en rechter constituenten en de keuze van de tussenklank. De resultaten van een produktie-experiment bevestigen dat de levendheid en de concreetheid van de linker constituent significant de keuze van de tussenklank beïnvloedt. Er zijn echter geen aanwijzingen voor een effect van de semantische klasse van het rechter lid. Een post-hoc analyse van de experimenten in hoofdstuk 2 bevestigt dit resultaat. Anderzijds is in alle post-hoc analyses het semantische effect van de linker constituent wel onafhankelijk van de vormeffecten van de rechter en linker constituentfamilies.

Hoofdstuk 5 onderzoekt het effect van linker en rechter constituentfamilies op de keuze van tussenklanken in nieuwe samenstellingen onder tijdsdruk, met het doel inzicht in het tijdsverloop van de selectie te krijgen. Een online-experiment waarin de snelheid van de keuze met behulp van knoppen wordt gemeten laat zien hoe belangrijk de distributie van tussenklanken op de keuze onder tijdsdruk is. Dit experiment repliceert de effecten op de keuze die de off-line experimenten in hoofdstuk 2 lieten zien. Bovendien laat het een analogisch effect van de linker constituentfamilie op de reactietijden zien: Een sterkere analogische steun gaat samen met kortere reactietijden. De rechter constituentfamilie beïnvloedt alleen de keuze, niet de reactietijd. Een simulatiestudie van het online-experiment met een implementatie van het interactieve activatiemodel van hoofdstuk 2 toont aan dat dit model zowel de effecten van de constituentfamilies op de keuze als op de reactietijden kan voorspellen.

Hoofdstuk 6 gaat over de vraag of constituentfamilies ook de keuze van tussenklanken in Duitse nominale samenstellingen kunnen beïnvloeden. Eerder onderzoek (Dressler, Libben, Stark, Pons, & Jarema, 2001) heeft hiervoor eerste aanwijzingen gegeven. De studie in hoofdstuk 6 repliceert het effect van de linker constituentfamilie op de drie meest voorkomende Duitse verbindingen *-s-*, *-(e)n-* en *-∅-* (*Seemann+s+Lied* 'zeemanslied', *Suppe+n+Topf* 'soeppan', *Haar+Farbe* 'haarkleur'). De rechter constituentfamilie speelt echter geen rol. Simulatiestudies van deze experimenten met TiMBL laten zien dat de selectie van Duitse tussenklanken zowel door de linker constituentfamilies als door eigenschappen van linker constituenten zoals rijm en geslacht beïnvloed worden. Er bestaat geen hiërarchie tussen deze factoren. Dit resultaat contrasteerd met de resultaten voor Neder-

landse tussenklanken. Hoofdstuk 6 eindigt met een schets van een psycholinguïstisch interactief activatiemodel dat de relevante factoren op de Duitse tussenklanken verenigt.

Hoofdstuk 7 is een studie over lexicale statistiek van woordvorming die de analytische werktuigen voor de statistische studie over tussenklanken in hoofdstuk 8 levert. Hoofdstuk 7 onderzoekt de over- en onderrepresentatie van complexe woorden in complexe woorden. Uit deze studie blijkt dat monomorfematische nomina vaker als constituenten in samenstellingen gebruikt worden dan men zou verwachten op basis van de proportie van monomorfematische nomina in de groep van alle Nederlandse samenstellingen. In tegenstelling tot monomorfematische constituenten zijn samenstellingen sterk ondergerepresenteerd als constituenten, en afgeleide nomina zijn zwak overgerepresenteerd. Verder laat deze studie zien dat de graad van overrepresentatie gecorreleerd is met de frequentie en lengte van de constituenten. Frequente en korte nomina zijn overgerepresenteerd, terwijl infrequente en lange nomina ondergerepresenteerd zijn. Voor afgeleide nomina geldt dat de graad van overrepresentatie ook gecorreleerd is met de produktiviteit van het suffix. Produktieve suffixen zijn ondergerepresenteerd, terwijl improduktieve suffixen overgerepresenteerd zijn. Dit patroon suggereert dat frequentere, kortere vormen met minder produktieve suffixen, die waarschijnlijk als geheel zijn opgeslagen in het mentale lexicon, makkelijker beschikbaar zijn voor verder woordvorming.

Hoofdstuk 8 onderzoekt de functie van tussenklanken als zij afgeleide linker constituenten volgen. Het gaat met name in op de functie van tussenklanken om 'closed suffixes' voor verder woordvorming te openen, zoals Aronoff en Fuhrhop (ter publicatie aangeboden) het voor Duitse tussenklanken voorstellen. Prototypische Nederlandse tussenklanken, zoals die door Booij en Baayen (in voorbereiding) gedefinierd worden, komen echter met en zonder tussenklanken voor. Bovendien komen ook 'non-closing' suffixen met tussenklanken voor. Daarom onderzoekt deze studie niet de combinatie van suffixen en tussenklanken, maar de graad van overrepresentatie van suffixen in samenstellingen als functie van de tussenklanken in deze samenstellingen. Suffixen blijken in samenstellingen die *-s-* of *-en-* bevatten in groter mate overgerepresenteerd te zijn dan in samenstellingen zonder tussenklank. Afgeleide woorden worden dus blijkbaar voor een gedeelte met behulp van tussenklanken 'geopend'. Deze functie lijkt de dominante functie van *-s-* te zijn. Voor de tussenklank *-en-* is er evidentie dat het ook het meervoud van het linker lid markeert (Schreuder, Neijt, Van der Weide & Baayen, 1999). Verder laat hoofdstuk 8 zien dat de correlaties tussen overrepresentatie en lengte of frequentie van

constituenten in samenstellingen die in hoofdstuk 7 geobserveerd werden alleen bestaan voor samenstellingen die geen tussenklank bevatten. Factoren die het voorkomen van tussenklanken bepalen, zoals de openingsfunctie, de constituentfamilie en de suffixfamilie, maskeren blijkbaar meer algemene factoren zoals frequentie en produktiviteit.

Hoofdstuk 9 presenteert naast een samenvatting van deze dissertatie ook een onderzoek naar de invloed van klemtoon op het voorkomen van de Nederlandse tussenklank *-en-*. Noch uit een statistisch studie van de samenstellingen in CELEX noch uit een simulatiestudie met TiMBL bleek enige evidentie dat het klemtoonpatroon van invloed zou zijn voor de keuze van de tussenklank.

De resultaten van deze dissertatie hebben belangrijke implicaties voor het lopend debat tussen de 'single-route'-benadering en de 'dual-route'-benadering van de verwerking van regelmatige en onregelmatige morfologische vormen. De 'dual-route'-benadering (bijvoorbeeld Pinker & Prince, 1991; Marcus, Brinkman, Clahsen, Wiese, & Pinker, 1995; Clahsen, 1999), die de traditionele linguïstische benadering representeert, neemt aan dat regelmatige en onregelmatige morfologische vormen door twee verschillend mechanismen verwerkt worden. Nieuwe regelmatige vormen zouden met behulp van regels gevormd worden, terwijl nieuwe onregelmatige vormen zouden naar analogie van reeds bestaande opgeslagen woorden gevormd worden. In deze benaderingen zijn regels het enige produktieve middel voor woordvorming. Nieuwe analogisch gevormde woorden zijn zeldzaam en exceptioneel.

In tegenstelling tot de 'dual-route'-benadering gaat de connectionistische 'single-route'-benadering (bijvoorbeeld Plunket & Juola, 1999; Rueckl, Mikolinski, Raveh, Miner & Mars, 1997; Seidenberg & Gonnerman, 2000) ervan uit dat regelmatige en onregelmatige gelede woorden door slechts een enkel mechanisme verwerkt worden. Nederlandse tussenklanken vormen een interessant proefterrein voor deze twee benaderingen omdat het gebruik van tussenklanken aan de éne kant voor het grootste deel met behulp van regels niet voorspeld kan worden en omdat het aan de andere kant volledig produktief is en een grote overeenkomst tussen sprekers vertoont. In een traditionele 'dual-route'-benadering zouden rond 70% van de samenstellingen in de experimenten in deze studie door exceptionele analogie verklaard moeten worden, een resultaat dat voor een produktief proces zeker niet bevredigend is. Zoals het onderzoek, beschreven in dit proefschrift, laat zien kunnen analogische modellen die hun voorspelling baseren op opgeslagen voorbeelden de keuze van tussenklanken met succes behandelen. Zoals in hoofdstuk 9 geargumenteerd wordt, zou hetzelfde resultaat ook met een connectio-

nistisch model bereikt kunnen worden. Omdat analogische modellen echter minder parameters bevatten en een interactief activatiemodel dat eenheden zoals morfemen en constituenten niet opgeeft transparanter is naar de taalstructuur toe, is in deze dissertatie voor de analogische benadering gekozen.

References

Aronoff, M and Fuhrhop, N submitted, Restricting suffix combinations in German and English Closing suffixes and the monosuffix constraint

Baayen, R H , Piepenbrock, R and Gulikers, L 1995, *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA

Booij, G and Van Santen, A 1995, *Morfologie De Woordstructuur van het Nederlands*, Amsterdam University Press, Amsterdam

Booij, G E 1996, Verbindingsklanken in samenstellingen en de nieuwe spellingregeling, *Nederlandse Taalkunde* **2**, 126–134

Booij, G E and Baayen, R H in preparation, Suffix order in Dutch

Clahsen, H 1999, Lexical entries and rules of language a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060

Daelemans, W , Zavrel, J , Van der Sloot, K and Van den Bosch, A 2000, TiMBL Tilburg Memory Based Learner Reference Guide Version 3 0, *Technical Report ILK 00-01*, Computational Linguistics Tilburg University

Dressler, W U , Libben, G , Stark, J , Pons, C and Jarema, G 2001, The processing of interfixed German compounds, *in* G E Booij and J Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp 185–220

Haeseryn, W , Romijn, K , Geerts, G , de Rooij, J and Van den Toorn, M 1997, *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff, Groningen

Marcus, G , Brinkman, U , Clahsen, H , Wiese, R and Pinker, S 1995, German inflection The exception that proves the rule, *Cognitive Psychology* **29**, 189–256

Mattens, W H M 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds), *De Nieuwe Taalgids* **7**, 333–343

Pinker, S and Prince, A 1991, Regular and irregular morphology and the psychological status of rules of grammar, *Proceedings of the 1991 meeting of the Berkeley Linguistics Society*

Plunkett, K and Juola, P 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490

Rueckl, J G , Mikolinski, M , Raveh, M , Miner, C S and Mars, F 1997, Morphological priming, fragment completion, and connectionist networks, *Journal of*

Memory and Language **36**(3), 382–405.

Schreuder, R., Neijt, A., Van der Weide, F. and Baayen, R. H.: 1998, Regular plurals in Dutch compounds: linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Skousen, R.: 2000, Analogical modeling and quantum computing, Los Alamos National Laboratory <http://arXiv.org>.

Toorn, M. C. v. d.: 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is, *De Nieuwe Taalgids* **74**, 197–205.

Toorn, M. C. v. d.: 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is, *De Nieuwe Taalgids* **74**, 547–552.

Toorn, M. v. d.: 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I, *De Nieuwe Taalgids* **75**(1), 24–33.

Toorn, M. v. d.: 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II, *De Nieuwe Taalgids* **75**(2), 153–160.

Curriculum Vitae

Andrea Krott was born in Aachen, Germany, on August 4, 1969. She studied Computational Linguistics and German Language and Literature at the University of Trier, Germany, and received her M.A. in 1995. From 1995 until 1997 she worked as a database manager at the Center for Lexical Information (CELEX) at the Max-Planck-Institut für Psycholinguistik, The Netherlands. From 1997 until 1998 she was a research fellow in the Linguistics Department at the Humboldt University of Berlin (Germany). In 1998, she returned to the Max-Planck-Institut für Psycholinguistik as the institute's corpus manager. In December 1998, she began her Ph.D in the PIONIER project 'The balance of storage and computation in the mental lexicon', funded by the Dutch Research Council (NWO), the Faculty of Arts of the University of Nijmegen, and the Max-Planck-Institut für Psycholinguistik. Currently she is a postdoctoral fellow in the Linguistics Department at the University of Alberta, Canada.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*

2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels Schiller*

3. Lexical access in the production of ellipsis and pronouns. *Bernadette Schmitt*

4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*

5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*

6. Gesture and speech production. *Jan-Peter de Ruiter*

7. Comparative intonational phonology: English and German. *Esther Grabe*

8. Finiteness in adult and child German. *Ingeborg Lasser*

9. Language input for word discovery. *Joost van de Weijer*

10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*

11. Producing past and plural inflections. *Dirk Janssen*

12. Valence and transitivity in Saliba, an Austronesian language of Papua New Guinea. *Anna Margetts*

13. From speech to words. *Arie van der Lugt*

14. Simple and complex verbs in Jaminjung: A study of event categorization in an Australian language. *Eva Schultze-Berndt*

15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*

16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*

17. Moving eyes and naming objects. *Femke van der Meulen*

18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*

ISBN 90-76203-11-3